Singapore teachers' classroom assessment: Preparing students for the "test of life," or a "life of tests"?

Author: Wei Ling Karen Lam

Persistent link: http://hdl.handle.net/2345/3804

This work is posted on eScholarship@BC,
Boston College University Libraries.

Boston College Electronic Thesis or Dissertation, 2014

Copyright is held by the author, with all rights reserved, unless otherwise noted.

Boston College
Lynch School of Education

Department of
Teacher Education, Special Education and Curriculum & Instruction

SINGAPORE TEACHERS' CLASSROOM ASSESSMENT: PREPARING STUDENTS FOR THE "TEST OF LIFE," OR A "LIFE OF TESTS"?

Dissertation
by

WEI LING KAREN LAM

Submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

2014

© Copyright Wei Ling Karen Lam

2014

Acknowledgements

A transnational village has sustained me through this PhD journey and I am truly grateful for the support everyone has given me.

First and foremost, I give thanks and praise to the almighty God for his providence and protection. He kept me safe and in good health as I ventured to Boston to start this journey.

I would like to thank my dissertation committee: my dissertation advisor, Dr. Andy Hargreaves, and my readers, Dr. Ina Mullis, Dr. Patrick McQuillan, and Dr. Dennis Shirley.

- I am greatly indebted to Dr. Andrew Hargreaves who, throughout my PhD journey, has been a source of inspiration and support. You opened my eyes to so many topics in education, and provided opportunity for a rich range of learning experiences. I have also learned so much from co-teaching with you and from working on the CODE project. Thank you for the detailed reading of the dissertation chapters, and for pushing me to think, and think more. Under your mentorship, I have grown more in this program than I ever imagined possible.
- To Dr. Ina Mullis, thank you for taking me as a graduate assistant at the TIMSS & PIRLS International Study Centre. The years I spent there have provided me with a very rich learning experience of international benchmarking studies. I am grateful for your detailed comments on the dissertation chapters, and for the consultation sessions in Singapore.
- To Dr. Patrick McQuillan who has taught me what it means to be a good researcher. I am thankful for the opportunity to work with you on the QCS project. The learning experience has been memorable and I will always be grateful for the advice you have given me.
- To Dr. Dennis Shirley, you guided me through the program in countless ways, from the FINAL project to writing a book review. I am inspired by your depth of knowledge, dedication to students, and love of learning. Thank you for your feedback on my drafts, and for taking time from your busy schedule while delivering speeches in Singapore to discuss the dissertation. Your support has meant more to me than I could possibly express.

I will always remember with warmth and gratitude my friends and colleagues at the TIMSS and PIRLS International Study Center. Your companionship and camaraderie created an inspiring and supportive environment in which to work and learn. I especially want to thank Dr. Michael Martin and Pierre Foy who gave me invaluable advice on the technical aspects of using the TIMSS data. I am also grateful to Corinna Preuschoff and Gabrielle Stanco Presley for their friendship and for taking me through the IDB Analyzer and SPSS procedures. Katie Trong Drucker, Alka Arora, Betty Poulos, Christine Hoage, Victoria Centurino and Jenn Bargioni have always been generous in lending a listening ear and in providing great advice.

To Dr. Andrew Hargreaves, Dr. Henry Braun and the CODE team (Maureen Hughes Ladieu, Kathryn Sallis, Alex Gurn, Matthew Welch, Lauren Chapman, Beth Morton, Youjin Lee and Adam Steiner), it is truly my honor and privilege to have had the chance to collaborate with you. Thanks also to Kristie Thayer who supported the project (and this dissertation) in so many ways.

To Dr. Marilyn Cochran-Smith, Dr. Patrick McQuillan, and the QCS team (Christine Power, Lisa De Souza, Cindy Jong-Halgin, and Kara Mitchell Viesca), thank you for an enriching first experience working on a university-based study, and for introducing me to the work of Newmann and Associates which inspired this study.

The LSOE community is caring, generous, and supportive, making the time spent there pass by very quickly. Here's to my doctoral cohort (Erin Hashimoto-Martell, Kate Ariemma Marin, Lyda Peters, Maureen Hughes Ladieu, Catherine Michener, Tracy Drysdale, and Laura Schalll-Leckrone) and all the other lovely people who have made a difference to my life and experiences at BC (Kathryn Sallis, Matthew Welch, Alex Gurn, Karen Terrell, Kathy Sillman, Jina Ro, Katherine Faircloth, Beth Morton, Apryl Holder, Liz Catron, Ken Jin Tan, Jenson Ong, and Deborah Wan).

My family has always been an important source of love, inspiration and support. Words cannot express how much you mean to me. I am deeply grateful to my parents, James and Juliana, for their care, love and encouragement all these years, for their belief in me, and for ensuring that I have had hot meals while working on the dissertation after work. My sister, Justina, and her family provided companionship during the holiday breaks. My beloved grandmothers who returned to the Lord just before this piece of work was completed, thank you for your unwavering love and support—this is for you!

To my dear friends in Singapore, thank you for believing in me, for your prayers, and for your words of encouragement. Thank you especially to Linda, Pei Ping, Puay Yin, Yim Ping, Shoo Yee, Noel, Angelia, May Yin, Pow Chew, Shao Ann, Soo Ben, Jeffrey, Hui Hua, Sylvia, Kathleen, and Fu Kuen. We were located on different continents but I cannot imagine completing the program without your friendship and encouragement, love and support.

I am also grateful to the Ministry of Education, Singapore for providing me with the opportunity and the funding to pursue a doctoral program. I wish to think especially Julie Tan who encouraged me to apply for the program, Tay Lai Ling who was a pillar of support, and Eugenia Tan who gave me the time and space needed to work on the dissertation.

This dissertation was made possible by the generosity of the teachers who participated in the study. I am heartened that many of you found that the deep reflections enriched your practices.

I would also like to thank Dr. Fred Newmann for answering my questions on the authentic intellectual work framework and for introducing the Centre for Authentic Intellectual Work website, to Dr. Val Klenowski for your encouragement and advice, and to Dr. Kim Hong Koh for sharing your experience rating teachers' classroom assessments.

I would be remiss if I did not acknowledge the assistance from the staff at the O'Neil Library who promptly delivered journal articles and book chapters which enabled work on this dissertation to proceed smoothly.

To my god children—Luke, Celeste and Emma—you are the next generation– this is for you.

And last, but not least, I owe an immeasurable debt to my best friend, supporter and husband, Peter. You have made significant sacrifices to enable me to reach this stage. Thank you for your love, patience, and kindness, especially for being the full time nurse when I fractured my leg during the last stage of this journey.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Singapore Teachers' Classroom Assessment:
Preparing students for the "test of life," or a "life of tests"?

Wei Ling Karen Lam
Dissertation Advisor: Dr. Andrew Hargreaves

In 2006, Singapore introduced the *Teach Less Learn More* (TLLM) movement to continue the systemic changes introduced under the *Thinking Schools Learning Nation* vision. A curricular initiative, TLLM had implications for classroom assessments, calling on teachers to focus on the process of learning, and to use more formative and qualitative assessing.

This dissertation examined the extent to which Singapore teachers' classroom assessment practices are aligned to the policy. It adopted mixed methods research to study teachers' assessment practices. Data culled from the Teacher Questionnaire used in the Trends in International Mathematics and Science Study provided the national pattern of assessment practices. Classroom practices were based on assessments contributed by eight teachers and from their interview comments. Classroom assessment practices were examined quantitatively using the Authentic Intellectual Work criteria (Newmann & Associates, 1996), and interpreted qualitatively using constructivist assessment (Shepard, 2000).

The findings suggest there was incremental change in the teachers' assessment practices. At the national and classroom levels, three patterns of assessment practices—change, variety, and persistence—emerged. Of the three, the pattern of persistence was the most dominant, indicating that most teachers continued to use assessment practices that the policy was discouraging. The prevalence of the pattern of persistence meant that teachers were more likely to focus on achievement rather than on learning. At the classroom level, the result of such assessment practices was that teachers did not always present students with challenging tasks.

There was a range of practices among the eight teachers. The extent to which the teachers' practices were aligned to the policy is the result of a complex interaction of policy, school, and classroom factors. Based on these findings, this dissertation suggests that to bring about fundamental change in classroom assessment practices, there needs to be greater macro policy coherence, a larger student role in the classroom, and more assessment leadership from principals.

CHAPTER 1: STUDENT LEARNING FOR THE 21ST CENTURY

Introduction

In the 21st century, technological improvements are rapidly changing the social, economic, political and cultural arenas of work, life and family. What knowledge, skills, and values must the young be equipped with in order to survive and thrive, and to contribute to a nation's progress? How should governments and the general public react and prepare themselves? And how should education respond to all this?

Within the United States, and in the countries belonging to the Organization for Economic Cooperation and Development, responses to the above have come from many interested groups (e.g., Common Core State Standards Initiative, 2010; European Parliament and Council of the European Union, 2006; North Central Regional Educational Laboratory & the Metiri Group, 2003; Organisation for Economic Co-operation and Develoment, 2005; Partnership for 21st Century Skills, 2011b) and individuals (e.g., Wagner, 2008) as to how to define, shape, and envision the needs of the future.

Perhaps the best known educational framework for preparing for the new world is the Partnership for 21st Century Skills, an American organization dedicated to promoting 21st century readiness for every child (Partnership for 21st Century Skills, 2011a, henceforth, the "Partnership"). The *Partnership* proposes a "holistic and systematic" framework for governments to "reconceptualize and reinvigorate public education" (Kay, 2010, p. xiv). The framework envisions students who are able to "design, evaluate, and manage their own work" through problem-solving, analyzing information, and synthesizing ideas to generate new knowledge (Darling-Hammond, 2010, pp. 33-34). These skills and dispositions embody the

features of the "knowledge economy, knowledge society" where "knowledge, creativity and invention are intrinsic to everything people do" (A. Hargreaves, 2003, p. 8).

The key components for the future curriculum advocated by the *Partnership* framework are core subjects; 21st century themes; learning and innovation skills; information, media, and technology skills; life and career skills, and 21st century education support systems (Partnership, 2011b). There are also corresponding calls to change assessment practices so that they match the envisioned curriculum framework. Groups like the Assessment and Teaching of 21st Century Skills (ATC21s.org, 2011) which are lead proponents of assessment reform point out that current assessment models fail to adequately measure the characteristics, skills, knowledge, and attitudes that are valued in the global economy. One argument is that external standardized tests provide inaccurate measures of student learning, given that they are detached from the curriculum, and typically can only measure facts and textbook knowledge because the test items are confined to a limited set of assessment formats (Eisner, 1991; Herman, Aschbacher, & Winters, 1992; Shepard, 1989; Torrance, 1995; Wiggins, 1992). Proponents of this position advocate for assessments that capture a larger set of skills, such as understanding and transfer of knowledge to new situations (T. L. Good & Brophy, 2008; Newmann & Associates, 1996; Perkins, 1999; Wiggins & McTighe, 2005), the ability to create rather than reproduce knowledge (Bryk, Nagaoka, & Newmann, 2000; Newmann & Associates, 1996), the capacity to engage in critical thinking and problem-solving (A. Hargreaves, 2003; Nitko & Brookhart, 2011), and the development of metacognitive skills (Shepard, 1989).

Since the knowledge economy values the production rather than the reproduction of knowledge, alternative modes of assessment to the current ubiquitous multiple-choice items are seen as being more congruent with the 21st century. Existing assessment types have been

critiqued for the way they assess the recall and regurgitation of discrete and disparate knowledge and facts. Comparatively, this new view conceives that assessments should be integral to teaching, such that teachers work alongside as students are completing assigned tasks, and give support, guidance, and feedback (e.g., Darling-Hammond, 2010; Shepard, 2001).

For students to be able to participate in the knowledge society, educators have to provide learning opportunities and assessments that nurture these skills. However, teachers lack, or do not believe they have the capability to design appropriate classroom assessment tasks (Erkens, 2009a). In fact, there is a gap between teachers' envisaged goals of student learning and the types of assessments they design or select to use (Bol, Stephenson, O'Connell, & Nunnery, 1998; Bol & Strage, 1996). While teachers want their students to develop higher-order skills such as interpreting information and thinking critically, the findings from empirical work indicate that a large proportion of their test items were based on recognition and recall of factual knowledge (Bol & Strage, 1996). This disconnect may be worrying because teachers are not assessing the type of higher-order thinking skills and abilities desired by the knowledge-based society.

Study Context

In Singapore, policy think tanks and government agencies have also embarked on a similar effort to prepare citizens for the 21st century. In 1997, the Ministry of Education (MOE) launched the *Thinking Schools Learning Nation* (TSLN) vision to transform the education system, and to prepare the young for the demands of the 21st century. *Thinking Schools* were given more autonomy to serve as "crucibles for questioning and searching, within and outside the classroom" (C. T. Goh, 1997, paragraph 22). Schools were places that "fire in … students a passion for learning" (C. T. Goh, 1997, paragraph 21). In a *Learning Nation*—the second half of

this education vision—learning is to transcend schools and educational institutions so that it occurs at every level of society. TSLN envisioned that Singaporeans

> must get away from the idea that it is only people at the top who should be thinking and the job of everyone else is to do as told. Instead we want to bring about a spirit of innovation, of learning by doing, of everyone each at his own level all the time asking how he can do his job better (C. T. Goh, 1997, paragraph 32).

The knowledge society as conceived by Singapore's leaders, involves participation in active citizenry, continuous learning, and grassroots-driven change. The focus on collective and individual intelligence and learning is the hallmark of a knowledge society (A. Hargreaves, 2003).

This policy decision for Singapore to embark on a new phase in education took place just as the results of the 1995 Third International Mathematics and Science Study (TIMSS) were released. Interestingly, Singapore students had out-performed their peers in other countries in mathematics in the third and fourth grades, while the secondary students were the highest achievers in mathematics and science in both the seventh and eighth grades. The students' performance was a tremendous success for Singapore, providing a measure of the progress in education made since the former British colony achieved independence in 1965. In spite of this stellar and sterling performance by Singapore's students, the launch of the new education vision and its related implementation strategies signaled that the country's leaders were carefully and strategically watching the regional and international political and economic arenas in order to steer the country into the 21st century (Ng, 2008).

The TSLN vision was to be realized through several strategies introduced and implemented gradually from 1997. One policy introduced to realize TSLN is an inspiring tagline, *Teach Less Learn More* (TLLM) which was put forth in 2004. As a significant lever in realizing

TSLN, *Teach Less Learn More* was mooted by the country's third Prime Minister, Lee Hsien Loong. He called on teachers to "teach less" so that students could "learn more" (H. L. Lee, 2004)

> I think we should cut down on some of this syllabus. It would mean less pressure on the kids, a bit less rote learning, more space for them to explore and discover their talents and also more space for the teachers to think, to reflect, to find ways to bring out the best in their students and to deliver quality results. We've got to teach less to our students so that they will learn more. Grades are important – don't forget to pass your exams – but grades are not the only thing in life and there are other things in life which we want to learn in school (H. L. Lee, 2004).

Ultimately, TLLM envisioned that the goal of education was to prepare students for the *test of life* rather than a *life of tests*, and eight years later, the Prime Minister reiterated this goal in his National Day Rally speech (see H. L. Lee, 2012).

The Prime Minister's inspiring call, which was the catalyst for the TLLM movement, has implications for curriculum, assessment and teaching. After all, how should teachers and schools interpret *teach less* and *learn more*? Policymakers explain TLLM as a return to the fundamentals of teaching, focusing on the *what*, *why*, and *how* of teaching, and calling for an improvement in the "quality of interaction" between teachers and students (MOE [Bluesky], 2005). Yet the very notion of "more" and "less" connotes visions of quantity (K. Tan, 2008). Another tension is evident from the Prime Minister's comment. On the one hand, he said that "grades are not the only thing in life;" in the same vein, TLLM reminds educators that the purpose of education is to provide young Singaporeans with "a quality of education that will prepare them for life, much more than prepare them for examinations" (Shanmugaratnam, 2005b). On the other hand, in the same quote, the Prime Minister called for teachers to "bring out the best in their students" in order to "deliver quality results." What does "quality results" refer to since, more often than not, *results* in Singapore are equated with test scores? For

educators in schools, understanding the nuances of these labels is critical as this affects how they interpret, incorporate, and implement the vision in schools as well as plan classroom learning events. To this end, one must ask: how do teachers teach and assess their students "qualitatively," as TLLM envisions?

In view of the above discussion of student learning and assessing higher-order cognitive skills, an important question arises: *What is learning?* Both TSLN and TLLM speak of "learning" – a *learning* nation, students *learn more* respectively, and as indicated in the Prime Minister's quote above, "other things in life we want to *learn* in school." None of these phrases uses the terms 'achievement' or 'performance.' Rather, under TSLN, learning is more than the acquisition of facts and content; indeed it is the "development of creative thinking and learning skills" (C. T. Goh, 1997). Like TSLN, *learning* under the TLLM banner is not about covering or learning the content. Instead, it is about meeting the "needs, interests and aspirations" of the learner, enabling the learner to be "passionate about learning," and teaching for "understanding of essential concepts and ideas" (MOE [Bluesky], 2005).

As with other (East) Asian students, the stereotypical view of Singapore students is that they are examination-smart, and highly skilled at rote learning and reproducing what they memorize when taking examinations. Western observers have attributed such behavior to the successful performance of Asian students in consecutive cycles of international achievement tests (Richards, 2004; D. A. Watkins & Biggs, 1996, 2001) and in high stakes national examinations. In Singapore, even under TSLN, the impact and importance of national examinations continue to prevail; high stakes national examinations at the end of primary 6 (Primary School Leaving Examination), secondary school (GCE 'O' level) and pre-university

(GCE 'A' level) are still used for placement at the next stage of education, and for certification.[1] Yet, in view of preparing students for the *test of life*, policy changes have been, and are being made to reduce the reliance on test scores and to access social and higher-order cognitive development.[2] Granting schools autonomy to accept up to 10 percent of students based on sporting and non-academic achievement is just one of the ways to reduce the weight of examination results when students progress from primary to secondary school. In a sense, it may be argued that there are paradoxes within the TSLN vision and TLLM tenets, and these tensions may impede the realization of the reform in the classroom.

While there are signposts indicating change at the policy level, what are the related changes in classroom practices at the ground level? In an important study of American schools over the period 1890 to 1980, Cuban (1984) reported that there had been little significant change in classroom practices, from the arrangement of the furniture to the pedagogy and assessment practices. Overall, he concluded that teacher-dominated strategies still prevailed in the classroom, despite advances in research on teaching. Some reasons for this persistence of traditional practices are the organization and structure of schools and classrooms, the culture of teaching, and teachers' personal and professional beliefs about the role of school, and classroom authority (Cuban, 1984).

With respect to Singapore's educational changes since 1997, this dissertation examines the extent to which teachers' classroom assessment practices are aligned to the TSLN vision. At the same time, if teachers are to produce "quality results" as mentioned in the Prime Minister's

[1] These are the three significant end-of-key stage national examinations that mark the end of primary, secondary, and pre-university education in Singapore. Students' performances in these assessments are used as placement at the next education stage and certification. More details can be found at http://www.moe.gov.sg/ and http://www.seab.gov.sg/

[2] The Prime Minister reiterated these aims in his address to the country on National Day 2013. See http://www.pmo.gov.sg/content/pmosite/mediacentre/speechesninterviews/primeminister/2013/August/prime-minister-lee-hsien-loong-s-national-day-rally-2013--speech.html#.UmaaJfmnpBo

speech, what do classroom assessment practices in Singapore schools look like? To signal a policy change, modifications to the national examinations were made with the introduction of group project work, and school-based and open-book assessments (Y. K. Tan, Chow, & Goh, 2008). Evidently, there is scope for more policy shifts because these new assessment modes have yet to be extended to all subjects and grade levels. Currently the new modes of assessment only apply to a small number of subjects at the GCE 'O' and 'A' level national examinations. Recently, the MOE announced that a new component focusing on 'investigation' would be introduced to the lower secondary history and geography syllabus in 2013. Instead of a formal examination, the 'investigation' component would assess students using a variety of assessment types like case studies (Channelnewsasia, 2012).

The key question that this dissertation explores is the nature and pattern of teachers' classroom assessment practices 15 years after the launch of TSLN and seven years after TLLM. Do the teachers' assessment tasks provide students with opportunities to apply higher-order thinking skills and to demonstrate their learning in different ways? Or are teachers still mimicking the format and modes used in national examinations? In line with the TSLN focus on deep learning and understanding, do teachers' assessment tasks provide avenues for students to demonstrate their learning rather than by showing their ability to recall facts. Can students show their ability to apply their learning to real life episodes or to transfer their learning beyond the classroom?

This dissertation study is guided by the following overarching research and supplementary questions:

Under an educational policy that emphasizes the preparation of students for "the test of life" instead of a "life of tests" (MOE [Bluesky], 2005), how do Singapore geography teachers elicit and enhance student learning through the ways they use classroom assessment?

1. From 1995 – 2011, what have been the patterns of Singapore teachers' classroom assessments?
    a. What forms of classroom assessments do Singapore science teachers report using in Secondary 2 (Grade 8) classrooms?
    b. How have the reported forms and patterns of classroom assessment changed over time?
    c. What are the associated patterns of student learning?

2. With respect to classroom assessment, how do Singapore geography teachers understand and use different forms of assessment in their teaching to address and enhance student learning?
    a. What does "assessment" mean to Singapore geography teachers?
    b. What is the nature and quality of classroom assessment that Singapore geography teachers create for their students?
    c. What is the nature and quality of work that students produce in response to teachers' classroom assessment?
    d. What is the relationship between the nature and quality of teachers' classroom assessment and student work?
    e. After implementing their classroom assessments, how do Singapore geography teachers make formative use of assessment data?

3. What factors influence the nature and quality of classroom assessments designed by Singapore geography teachers in response to the *Thinking Schools, Learning Nation* vision?

Significance of the study

This study is significant for several reasons. First, it contributes to the research on Authentic Intellectual Work (AIW). Specifically, it uses Newmann and Associates' (1996) AIW criteria as indicators of the higher-order thinking skills envisioned by TSLN to examine geography classroom assessments—teacher assessment in this subject is not extensively examined, nor is it examined using the AIW criteria. In particular, this dissertation provides an in-depth study of the types of assessments eight geography teachers present to their students, and seeks to understand how and why the teachers assess the subject so differently, despite assessment objectives presented in the syllabus. Furthermore, the empirical work on AIW has focused on rating the quality of teacher assessment using quantitative methods. This study

builds on this field by including interviews with teachers to obtain a deeper understanding of their conceptions of and rationale for the assessment tasks that they assign their students.

Second, this dissertation examines student learning through the use of multiple indicators. In particular, it looks at student learning by analyzing the scores in the cognitive and content domains in five cycles of TIMSS, and by qualitatively examining student work completed in response to the assessment tasks designed by their teachers. More specifically, this dissertation uses student achievement scores in consecutive cycles of TIMSS data (cycles 1995, 1999, 2003, 2007, 2011) as an indicator of student learning. In addition to examining the overall achievement scores in TIMSS, this study also uses the different cognitive domains—knowing, applying, and reasoning—as provided in the TIMSS database to examine student learning. In this way, student learning is examined in terms of both their achievement in the content and cognitive domains. Based on this, this dissertation's working hypothesis is that with the deepening of TSLN practices in the last 15 years, in particular with the emphasis on thinking skills, Singapore students' scores in the *reasoning* and *applying* domains of the TIMSS assessments should show a trajectory of increase over time. This study extends the concept of student learning by analyzing how teachers use formative assessment strategies, in particular feedback, to help students progress from the current learning status toward the intended goals.

Third, this study contributes to the research on formative assessment. The review of over fifty pieces of empirical work on formative assessment (in Chapter 2) indicates a preponderance of work focusing on teachers' beliefs and the impact of their practices on student achievement. There are fewer studies examining the way teachers apply formative assessment after they analyze and interpret student work, especially in a country like Singapore. This study adds to the

corpus of work on formative assessment by examining Singapore teachers' use of formative assessment to enhance student learning.

Fourth, there is a rich body of empirical and conceptual work on authentic intellectual work and formative assessment. However, based on the review of the literature on these two concepts, few studies link the two together. To this end, this study combines the two concepts in order to examine the quality and nature of student learning and teacher assessment in eight Singapore teachers' classrooms. This study integrates the two concepts sequentially: first by using authentic intellectual work as an indicator for the skills and capabilities consistent with those envisioned in TSLN as being critical for life beyond school in the 21st century, and second, by exploring how Singapore teachers use formative assessment to enhance student learning. In so doing, this dissertation suggests that the combined analyses of interviews with teachers, of teacher assessment and of student work provide a means by which to understand the conditions that contribute to the assessing of higher-order thinking skills and other 21st century capabilities and dispositions in Singapore under a new policy vision.

Theoretical and conceptual framework

This dissertation draws on constructivist learning theories as a theoretical lens to interpret and analyze formative assessment and authentic assessment practices, and to examine how these interact and play out in Singapore teachers' classroom assessment practices. With respect to authentic assessment, this dissertation applies the AIW criteria (Newmann & Associates, 1996) as indicators of the type of 21st century learning envisaged by TSLN and TLLM. Second, formative assessment practices as conceived and popularized by Black and Wiliam (1998a) are the means and processes of realizing the vision, and of helping students *learn more* within the TLLM tagline.

Current interest in constructivist learning theories emerges from research in cognitive science which conceives of learners as active participants in the learning process who construct learning by interpreting and incorporating new experiences and knowledge into their prior knowledge and learning (T. L. Good & Brophy, 2008; Graue, 1993; Perkins, 1999; Resnick, 1987b, 1989; Shepard, 2000). In response to this, the role of the teacher changes from the dispenser to the facilitator of knowledge (Graue, 1993; Schiro, 2008). Because students take time to interpret and incorporate new and novel information and experiences, learning is seen as a process that develops over time. Assessment that is aligned to this view of learning and teaching, documents change over time and does not merely provide a moment-in-time report. To this end, formative assessment is an important practice in helping students move from their current state of learning to the desired level (Black & Wiliam, 1998b; Sadler, 1989). This is achieved by teachers using formative strategies (see Black, Harrison, Lee, Marshall, & Wiliam, 2003b).

A key tenet of constructivist learning involves students building on their prior knowledge when they are exposed to new information and experiences. The corresponding assessment practices should provide opportunities for students to respond to and apply what they know to new and different situations. Such assessment is "ecological" (Biggs, 1996a) because it requires the application of current knowledge to contexts that are "authentic" or ecologically valid. Scholars like Archbald and Newmann (1988), Newmann and Associates (1996), and Wiggins (1989) use the term "authentic" assessment, arguing that the contexts need to mirror the behavior of experts in the real world. Despite the differences in the concepts of "authentic," there is agreement that assessments within constructivist theories of learning should be carried out over a period of time so that change is developmental (Biggs, 1996a; Wiggins, 1989), should require

the construction of knowledge (Newmann & Associates, 1996), and should focus on higher-order thinking and understanding as well as the synthesis and integration of knowledge (Newmann & Associates, 1996; Shepard, 2000).

This dissertation suggests that the tenets and spirit of the TSLN-TLLM reforms are consistent with the essential features and characteristics of constructivist learning theories. Singapore's efforts at educational change as manifested in the TSLN vision, and in TLLM's call on teachers *teaching less* so that students can *learn more* are intended to improve the "quality of interaction" between teachers and students (MOE [Bluesky], 2005). Chapter 2 provides a comparison of the similarities between TSLN-TLLM and constructivist learning theories.

Research design

This study uses a mixed methods research design to examine the realization of an educational policy that urges Singapore teachers to use more "formative and qualitative" assessments (MOE [Bluesky], 2005). Embracing a combination of quantitative and qualitative research programs to study educational policy provides insight into policy implementation and policy consequences that would otherwise be unattainable if just one method or approach were used (Desimone, 2009; Luke & Hogan, 2006; M. L. Smith, 2006).

Mixed methods research is underpinned by pragmatism, adopts problem-centered approaches, embraces pluralistic world views and perspectives, and is oriented toward real-world practices (Creswell & Plano Clark, 2007). Mixed methods approaches are appropriate for this dissertation because the study comprises multiple research questions that drive the research methodology. Furthermore, this methodology supports the research questions proposed for this study that other methodologies cannot (Teddlie & Tashakkori, 2003), and allows for a more

insightful and comprehensive understanding through the combination of quantitative and qualitative methods (Greene, 2001).

The use of mixed methods approaches enables the combination of macro and micro data, as well as primary and secondary data. Specifically, data culled from the secondary analyses of TIMSS provides a *macro* picture of teachers' assessment practices over time. This phase of the study uses secondary data. Subsequently, to obtain deeper insight of the *micro* level (or classroom assessment practices), this dissertation uses primary data, collected from eight teachers participating in the study. This phase provides a picture of current assessment practices. Together, the primary and secondary data provide national (macro) and classroom (micro) patterns to examine Singapore teachers' classroom assessment practices conducted and implemented to realize the TSLN vision.

This dissertation examines the patterns of Singapore geography teachers' classroom assessment since the launch of the milestone TSLN vision in 1997 and its follow-up initiative, TLLM in 2004. It uses a mixed methods approach to combine the analyses of multiple data sets. First, the TIMSS contextual questionnaires over five cycles are used to analyze the macro (national) pattern of teachers' assessment practices. Second, curriculum and teaching documents, interviews with teachers, and teachers' classroom assessment tasks and student work are examined to understand the enactment of this policy and its translation in the classroom. The teacher self-report contextual data from five cycles of TIMSS (1995, 1999, 2003, 2007 and 2011) provide indicators of Singapore teachers' classroom assessment over the 16 years that parallel the existence of the TLLM vision, while the interviews, assessment tasks and completed student work provide indicators of current assessment practices.

Singaporean students' achievement in the content and cognitive domains in the five TIMSS cycles provides an indication of learning over time. The hypothesis is that in each successive TIMSS cycle, there should be an increase Singapore students' achievement in the *reasoning* and *applying* domains, if the intent of TSLN-TLLM is being practiced or implemented. As another indicator of Singapore's students learning, this dissertation analyzes teacher assessments and student work; the former serves as a proxy for the quality of assessment tasks presented to students, and the latter as an alternative indicator of the quality of student learning. To further examine how teachers enable Singapore students to *learn more*, there will be in-depth interviews with teachers to examine how they interpret student work, and consequently make pedagogical and curricular decisions to provide formative feedback to students, enabling them to move from their current level of learning towards the intended learning goals and levels.

Organization of the study

This dissertation consists of seven chapters. The first chapter introduces the study, its context, and its contribution to the field. It provides the theoretical and conceptual perspectives that underpin the study, as well as a short overview of the research methods. Chapter 2 comprises a review of the theoretical and empirical literature on classroom assessment, authentic assessment, and formative assessment. Through the review, gaps in the extant work are identified to direct the research questions and methods used in this study. Chapter 3 presents and explains the research methodologies used in this dissertation. It also outlines the data collection, analyses, and interpretation procedures. The data analyses are presented in Chapters 4, 5, and 6. Chapter 4 focuses on the analysis and discussion of the *macro* data culled from documentary and secondary analyses of five cycles of TIMSS and thereby, paints a national picture of teachers' assessment practices over time (from 1995 to 2011). Both Chapters 5 and 6 present the analyses

and discussion of the classroom or *micro* data obtained from the interviews conducted with teachers over a twenty-week period. The set of primary data is presented in two chapters, with Chapter 5 focusing on the quantitative analyses, and Chapter 6 emphasizing the qualitative analyses. For each of the chapters, I also discuss the research and analysis procedures. Finally, Chapter 7 offers the meta-inferences based on the analysis, and suggests implications for policy, research and teacher education, and practice in schools.

CHAPTER 2: LITERATURE REVIEW

This chapter begins with a discussion of constructivist learning theories, the theoretical perspective that drives this dissertation. Using this theoretical frame, this dissertation examines the relationship between constructivist learning theories, assessment, and student learning, and the embodiment of these ideas in Fred Newmann and Associates' (1996) conception of *authentic intellectual work* (AIW) and Black and Wiliam's (1998a) notion of *formative assessment*. Next, the chapter presents a review of the conceptual and empirical literature on *authentic intellectual work* and *formative assessment*. The chapter concludes with a summary of the extant empirical research on authentic intellectual work and formative assessment, and identifies the gap in the current body of research that this dissertation will contribute towards filling.

Theoretical perspective

One of the most important words in education is *learning* (Cole, 1990). Yet, it is difficult to define and to observe what learning is or when it has taken place (H. H. Marshall, 1992b). While tests or assessments are used as indicators of learning, they are limited in the ways they measure what students have learned.

The word *learn* features prominently in Singapore's *Thinking Schools Learning Nation* (TSLN) vision, and in one of its implementation approaches, *Teach Less Learn More* (TLLM). What then does *learn* mean? The vernacular definition of *learn* points to the following: find out about something, obtain knowledge and skill after deep study or training, experience a change in behavior or attitude after being exposed to the desired way, study so as to be able to repeat something (e.g., a poem) (Collins Cobuild Advanced Learner's English Dictionary, 2005). These definitions present two aspects of the word *learn*: first, that it is a process, and second, that there is an outcome or an ability to demonstrate the result of the process. Scholars like Chris Watkins

(2011) present a more complex view of learning that involves making sense of information (Tinzmann, Jones, & Pierce, 1991): (1) learning is being taught; (2) learning is individual sense-making; and (3) learning is building knowledge as part of doing things with others (C. Watkins, 2011). The next section discusses different views of learning, and suggests that how *learning* is presented within different learning theories illustrates its complexity, and how it may be effectively achieved (Ertmer & Newby, 1993).

Dominant views of learning

Learning theories are useful in providing educators with information on the various pedagogies and techniques available (Ertmer & Newby, 1993). When selecting teaching strategies to meet particular educational objectives, educators need to be aware of the choices available. Learning theories are the foundation to facilitate reasonable decisions regarding the selection of strategies to use (Ertmer & Newby, 1993). Without an appreciation of the underlying theoretical basis for education change and activities, educators may not be able to enact or translate the change effectively (Black & Wiliam, 2012b). However, learning theorists do not explain or articulate explicitly how to assess learning outcomes within their models (James, 2006; James & Lewis, 2012). This omission may be due to an inadequate theoretical base for some types of assessment practices, or may suggest that the evolution of newer learning theories has yet to be matched by conceptual developments in the assessment (James, 2006). As a result, there is a disconnect between theories of instruction and assessment and often, there are "only tenuous or partial relationships with current understanding of learning" (James, 2006, pp. 47-48). Elizabeth Graue (1993, p. 291) describes this misalignment as assessment and teaching being "conceived as curiously separate in both time and purpose." Classroom instruction seems

to adhere to newer learning theories while assessment continues to be aligned with older, more traditional psychometric paradigms (Graue, 1993; James, 2006; Shepard, 2001).

As this dissertation draws largely on constructivism, its key features as well as its assumptions and philosophies related to learning and assessment are presented in greater detail in the following sections. The discussion of constructivist learning theories is presented in contrast to behaviorist theories. This is because in the literature, these two philosophies are often portrayed as two diametric poles of a continuum, each embodying different epistemological positions, emanating from different views of teaching and learning, and ultimately creating different implications for assessment.

With reference to these two learning theories, this dissertation suggests that Singapore's TLLM rhetoric is calling on educators to increase their use of constructivist methods to help students *learn* more. I also suggest that the policy is not advocating a pendulum-like swing from one paradigm to the other, but is calling on educators to adopt and use a wider repertoire of pedagogies so that they can meet the learning needs of diverse groups of students.

Behaviorist and constructivist theories of learning and assessment

Constructivist learning theories are a family of theories (Biggs, 1995) that include social constructivist, cognitive, and sociocultural constructivist theories (T. L. Good & Brophy, 2008). According to James (2006), different nomenclatures are used for constructivist theories; in the USA, they are known as cognitive or situated learning, while in the European literature, they are referred to as sociocultural learning. The growing popularity of constructivist learning theories is due to research from the field of cognitive science which has provided substantial insight into how humans think and learn (see for example, Bransford, Brown, & Cocking, 2000; Pellegrino, Chudowsky, & Glaser, 2001).

Constructivist learning theories are believed to have deeper historical roots than behaviorist theories, although the latter have dominated the greater part of the 19th and 20th centuries. The roots of constructivist tenets in teaching and learning can be traced back to John Dewey, Jean Piaget, and, some academics suggest, even as far back as Socrates (T. L. Good & Brophy, 2008). In the 20th century, Vygotsky was a leading influence among constructivists (Gipps, 1999). Comparatively, behaviorist learning theories are believed to have gained ascendency only from the 1930s (Kliebard, 2004; Schiro, 2008).

The literature is replete with metaphors and labels for these two sets of learning theories, including: (1) acquisition and participation metaphors of learning (Sfard, 1998); (2) quantitative and qualitative traditions of learning and assessment (Biggs, 1992, 1995, 1996a); (3) 20th century dominant paradigm and "emergent constructivist paradigm" (Shepard, 2000, 2001) and (4) traditional and constructivist models (Gipps, 2002). Regardless of the nomenclature, the different metaphors and labels for the theories have similar conceptions of learning, of the roles of teachers and students in teaching and learning, and of the ensuing assessment practices.

Behaviorist learning theories are associated with the work of Pavlov, Skinner, and Thorndike (James, 2006). The dominance of these theories today is evident in the ubiquitous application of standards-based curriculum and prescriptive teaching packages. Within the behaviorist tradition, knowledge is a well-defined body of information (Gipps, 2002), and comprises basic atomistic pieces of information or facts that can be accumulated (Shepard, 2001). For this reason, Sfard (1998) calls this the "acquisition" metaphor of learning, Biggs (1996a) uses the term the "quantitative" tradition of learning, Wolf, Bixby, Glenn and Gardner (1991) label this "scalar learning," and Freire (2000) adopts the term "banking education." Together,

these terms for behaviorist theories suggest that knowledge and learning are seen as processes of progressively aggregating and accumulating information.

Within the behaviorist tradition, learning is perceived as the passive acquisition of facts, concepts, and skills, typically achieved through guided, routine or mimetic practices (Brooks & Brooks, 1993), sometimes encouraged by incentives or punishments (H. H. Marshall, 1992b). Learning in the behaviorist tradition is manifested by the modification of behavior. To this effect, frequent testing is believed to be necessary to ensure mastery of the hierarchical skill sets and knowledge (James, 2006). Given the stair-like nature of knowledge, students are only able to advance to more complex tasks and knowledge after they have mastered the foundation (Cole, 1990; James, 2006; Resnick, 1989; Tinzmann, et al., 1991). For this reason, this learning theory is also known as the "basic skills model" (Tinzmann, et al., 1991).

The typical assessment within the behaviorist tradition or what James (2008, p. 21) calls "first generation assessment practice" has specific features: it assesses discrete pieces of competence, facts or knowledge (Cole, 1990); it is a timed assignment; and it interprets students' mastery as binary "correct" or "incorrect" responses (James, 2006, p. 54). These binary responses are aggregated to provide a total score that is a representation of competence (Biggs, 1996a). To achieve the desired behavior or level of competency, students who do not perform as expected are subject to remedial action, often involving more practice of the incorrect items (James, 2006).

The criticism of behaviorist theories of learning is their incongruity with current knowledge about human learning. In particular, the focus on discrete bits of knowledge has been criticized for diluting the curriculum and for focusing only on low level competencies (Shepard, 2001). The overreliance on assessment modes such as multiple choice and short-answer

constructed response items has also been judged unfavorably. Such assessment modes, Eisner (1991) bluntly argues, dumb down both teachers' professionalism and student learning because neither is required to make any sort of judgment

> When there are five alternatives for the student to select and only one correct response among the five, scoring can be handled by an optical scanner; no one needs to exercise any judgment whatsoever. … Although such procedures are efficient, they prohibit test makers from asking the kinds of questions that do not fit a predetermined correct answer (Eisner, 1991, p. 173).

Since the late 1980s, an "emergent constructivist paradigm" (Shepard, 2000, p. 5) of curriculum, teaching, and assessment has risen to challenge the dominance of behaviorist learning theories. Seen as a response to the growing dissatisfaction with behaviorist learning theories, constructivist learning theories are antithetical to the efficiency model (Shepard, 2000). Constructivist theories are closely related to the concept of the mind, and conceive of learning as an active process of sense making (Biggs, 1996a; T. L. Good & Brophy, 2008; Shepard, 2000). The family of constructivist theories shares similar principles: students build on prior knowledge, actively construct knowledge, mediate and make sense of what they learn by relating new experiences to what they already know (T. L. Good & Brophy, 2008).

Compared to behaviorist theories which assign curriculum based on students' present or anticipated station in life (Shepard, 2000), constructivist learning theories embrace an inclusive view of education in which 'All children can learn' (Shepard, 2000, 2001). Within this tradition, knowledge is no longer segmented into atomistic parts or silos, but is horizontally integrated into other topics and subjects, as well as vertically interconnected with prior and ensuing learning (Biggs, 1996a). As a result, learning is a gradual process during which students interpret, incorporate, and cumulate new information, knowledge, and facts by building on previous knowledge (Biggs, 1996a; Ertmer & Newby, 1993). Since understanding evolves progressively

with new learning, Biggs (1996a) refers to this as the "qualitative" tradition of teaching. This is because there are qualitative changes within the nature of "what is learned and how it is structured" (Biggs, 1996a, p. 3). To this end, learning is not about the reproduction or re-presentation of knowledge and facts, but is also the creation of new knowledge or meaning (Ertmer & Newby, 1993).

In comparison to behaviorist theories, constructivist theories exact higher demands on students, requiring them to relate prior conceptions and background experiences to new situations and contexts (T. L. Good & Brophy, 2008). While behaviorist theories are criticized for their focus on basic skills, scholars (e.g., Cole, 1990) purport that constructivist theories seek to elicit *higher-order thinking skills* or what Norris and Ennis (1989) term "critical thinking." These higher-order thinking skills require an expanded use of the mind such that learners have to interpret, analyze or manipulate information. These traits of higher-order thinking are evident when learners are required to solve a "fuzzy" problem that cannot be explained by applying routine recall of previously acquired facts and knowledge (Newmann, 1992).

Unlike in the behaviorist tradition, the teacher's role is not to transmit knowledge to passive students, but to support them in their efforts to construct understandings that gradually become more sophisticated. As a result, learning requires closer interactions between learners and those around them (Biggs, 1996a; James, 2006; H. H. Marshall, 1992b; Tinzmann, et al., 1991). One essential tenet of constructivist learning theories is that students are lively participants during the learning process. They "develop new knowledge through a process of active construction" (T. L. Good & Brophy, 2008, p. 337) and are in "the constant flux of doing" (Sfard, 1998, p. 6).

Given that learners construct and build on prior knowledge over time, assessment that is aligned to this learning theory is dedicated to "charting longitudinal growth" (Biggs, 1996a, p. 4). For this reason, formative assessment is a critical piece in constructivist learning theories because of its developmental emphasis and its focus on helping students understand their misconceptions, and thereby enabling them to move from their current status towards higher levels of attainment.

Since a key feature of learning in constructivism involves students interpreting and incorporating prior experience and knowledge into new and novel learning, the corresponding assessment practices should enable students to apply what they know to new situations (Newmann, Smith, Allensworth, & Bryk, 2001) as well as to demonstrate their growth over time. Scholars propose two assessment frameworks for constructivist theories of learning. First, assessments should be "ecological" because students are required to apply current knowledge to contexts that are 'authentic' or ecologically valid (Biggs, 1996a). Others like Archbald and Newmann (1988), Newmann and Associates (1996), and Wiggins (1989) prefer the term 'authentic' assessments, meaning that the assessments mirror contexts, situations, or the behavior of experts in the real world. Second, assessment should serve a developmental function, which is to discover where students currently are in their level of understanding or competence, and to help them progress to the next more advanced level. Unlike standardized tests in the behaviorist tradition which are implemented in unrealistic timeframes, constructivist assessments are to be conducted over a period of time so that developmental change may be documented (Biggs, 1996a; Wiggins, 1989). Other features of constructivist assessment include the construction of knowledge (Newmann & Associates, 1996), focus on higher-order thinking and understanding, as well as the synthesis and integration of knowledge (Newmann & Associates, 1996; Shepard,

2000). However, these conceptions of assessment are nascent. The "assessment technologies" (James, 2006, p. 6) proposed for constructivist assessments are still in their infancy, and as they are not undergirded by a theory of learning, are less valued than objective, standardized assessments.

Under the umbrella of constructivist learning theories, there are social, sociocultural, and cognitive constructivist theories. Social and cognitive constructivism differ in terms of the nature and influence of the social world during the knowledge construction process (H. H. Marshall, 1992b). Social constructivism situates thinking and learning within social contexts while cognitive constructivism views learning and thinking as taking place within the individual's mind (James, 2006; H. H. Marshall, 1992b). In social constructivist learning, social interaction plays a significant role in the way knowledge and meanings are constructed and structured (H. H. Marshall, 1992b). As a result, learners interact dynamically with their social context as they continually make sense of new experiences. The social contexts include interactions between teachers and students (H. H. Marshall, 1992b). Some scholars distinguish social and socio-cultural constructivist theories (T. L. Good & Brophy, 2008) while others (e.g., H. H. Marshall, 1992a) appear to combine the two categories. The interaction with other individuals is an essential feature of both social and sociocultural constructivism. The difference is that social constructivism involves "sustained discourse" and communication (T. L. Good & Brophy, 2008, p. 340) while sociocultural constructivism involves "enculturation" within a "community of practice" that ultimately enables novices (learners) to learn from mentors (teachers) (T. L. Good & Brophy, 2008, p. 342).

Cognitive constructivism is concerned with how the individual learner makes sense of information and knowledge. It is usually perceived as a theory of cognitive development,

focusing on how individuals construct meaning as they organize and relate concepts and information within their existing memory or knowledge base (Ertmer & Newby, 1993), and developmental models (James, 2006). In this theory, the focus is placed on how learners acquire knowledge (Ertmer & Newby, 1993) rather than on what learners do with the knowledge. The leading scholars in this area are Lev Vygotsky because of his conception of the *zone of proximal development* (ZPD) (H. H. Marshall, 1992b), Noam Chomsky, and Jerome Bruner (James, 2006). Drawing on the ZPD, cognitive constructivists view the teacher as helping learners arrive at expert understanding of more complex concepts through the use of processing strategies such as deductive and inductive reasoning to solve problems (Ertmer & Newby, 1993; James, 2006, 2008).

Within cognitive constructivism, formative assessment plays a strategic role (James, 2006) in moving learners from the novice to the expert level (James, 2008). Feedback as a formative assessment strategy is important because it serves as new knowledge or information that learners can actively integrate into their existing cognitive framework (Ertmer & Newby, 1993). Given the nature of such tasks, assessments within the cognitive constructivist tradition generally do not have one specific correct response, require more time to complete, and are based on specified criteria (James, 2008).

Another branch within the family of constructivist learning theories is socio-cultural constructivist theories. The theoretical origin of socio-cultural constructivist theories, estimated to be around the early 20th century, is believed to pre-date both behaviorist and cognitive constructivism theories (Bredo, 1997, in James, 2008). The leading scholars in this area include William James, John Dewey, and George Herbert Mead and later, Lev Vygotsky, and Yrjö Engeström (James, 2008), and their philosophical underpinnings are drawn from social theory,

sociology, anthropology, and psychology (James, 2006). The distinctive feature of socio-cultural learning theory is that learning and thinking take place via actions which change the situation. In turn, this situation alters the learning; thus, learning and the situation are constantly interacting (James, 2006).

In comparison to cognitive constructivism, sociocultural constructivists believe that learning is a mediated and social activity in which cultural tools, artifacts, and collaboration play important roles to help learners build knowledge and further their thinking (James, 2006, 2008; Shepard, 2001). As learning occurs in a communal setting, no one owns what is learned; rather, learning is distributed within the community. When individuals create or internalize new knowledge, they communicate it externally to those around them, who in turn also incorporate this knowledge. To this end, "knowledge is created and shared in expansive learning cycles" (James, 2008, p. 30). Such learning takes place most efficiently within a *zone of proximal development* (T. L. Good & Brophy, 2008; Vygotsky, 1978), and consequently, teachers need to create a learning environment in which learners are encouraged to think about and respond to authentic tasks that are above their existing ability level (but within the ZPD). In designing the learning, teachers create activities that learners are able to complete on their own, although at the same time, some may need another person (e.g., teacher or peer) to provide assistance. Thus, the teacher needs to be aware of what scaffolding to provide and when to remove this support once the learner demonstrates that he or she is able to manage alone.

Learning within the community is based on the apprenticeship or "on-the-job" training metaphor (T. L. Good & Brophy, 2008, p. 342). Novices new to the community begin learning by observing, listening and working on entry-level activities at the periphery of this community

before being promoted to activities at the center as they gain expertise. In the process, novices learn through "legitimate peripheral participation" (Lave & Wenger, 1991, p. 29).

The implications for assessment, however, are less defined than those concerning teaching, and in particular, the implications of sociocultural theory for assessment have yet to be clearly articulated (James, 2008; James & Lewis, 2012). This may be attributed to theorists prioritizing learning over assessment. Consequently, more effort is needed to define and delineate the alignment between sociocultural learning, teaching, and assessment (James, 2008; Torrance & Pryor, 1998). Ways to bring about congruence between learning, teaching, and assessment within sociocultural theory include:

- situated assessment taking place concurrently with learning;
- assessment done by the community rather than by external parties;
- assessment of group learning in addition to individual learning; and
- use of multiple assessment approaches to capture, document, and report learning outcomes (James, 2008).

Overall, assessment needs to change from "static" to "dynamic" (Gipps, 2002) because learners take time to build on existing knowledge. This means that rather than capturing a snapshot of learning, usually at the end of a unit or a school year, assessments could take place more frequently, involve tasks that require completion over a period of time (James, 2008), and include teacher (or expert) feedback for improvement. To work within the ZPD, students should have the opportunity to act on the feedback to improve their work (Sadler, 1989).

Summary. While mindful of the theoretical, philosophical, and conceptual differences between behaviorist and constructivist theories of teaching and learning, most scholars do not reject one for the other. Rather, scholars (e.g., Cole, 1990; T. L. Good & Brophy, 2008; James, 2006; Sfard, 1998; Stobart, 2008) caution the overreliance on one over the other, and suggest that educators may be justified in combining approaches, depending on their learning and curricular

intentions. Since teaching and learning are complex, the overlap of theories is inevitable (James, 2006). Furthermore, Good and Brophy (2008) recognize that "more responsible" scholars would agree that a complete educational experience includes teaching that presents information, as well as use constructivist methodologies (Sfard, 1998). The balance often depends on the goals of the particular lesson, such that the presentation of information may be more efficient for teaching canonical knowledge and basic skills while constructivist methods may be more appropriate when developing skills and processes (T. L. Good & Brophy, 2008; James, 2006).

"Emergent constructivist" paradigm of teaching, learning, and assessment. Since the 1990s, the emphasis on the use of high-stakes test scores as indicators of accountability and of the quality of education has created an "assessment society" that values "numbers, grades, targets and league tables" (Broadfoot & Black, 2004, p. 19). The high premium placed on assessment data has a backwash: rather than teaching and learning driving assessment, the converse is true – assessment drives teaching (Dahlin, Watkins, & Ekholm, 2001) and educational goals (Camp, 1992). The consequence is that in the classroom, assessment shapes teaching and learning (Gulikers, Bastiens, & Krischner, 2004; D. A. Watkins & Biggs, 2001), such that teachers' classroom assessment mimics the design, form, and structure of external assessments in order to prepare students to successfully accomplish these assessments, rather than tailoring these to meet teaching and learning needs. This results in a dissonance between teaching, learning and assessment because teaching is aligned to constructivist ideas whereas assessment continues to use principles from the behaviorist paradigm. In response, Lorrie Shepard (2000) proffers an "emergent constructivist" curriculum with teaching and learning theories, and classroom assessment framework (See Figure 2.1) that combines key features from cognitive, constructivist, and sociocultural theories in order to define and delineate a curriculum

vision and its corresponding teaching and assessment practices. The principles envisaged in the framework, Shepard acknowledges, are antithetical to those from the behaviorist tradition. In 2000, Shepard called this an "emergent" framework because at that time, it had yet to evolve theoretically, and the practices had yet to be adopted. Even today, the implications of new learning theories on assessment are yet to be clearly defined (James, 2008; James & Lewis, 2012). To this end, while scholars have communicated their conceptions of constructivist assessments, there are issues that warrant deeper debates and examination.

Figure 2.1.
Emergent constructivist paradigm (adapted from Shepard, 2000, p.8).

(a) Reformed Vision of Curriculum

- All students can learn.
- Challenging subject matter aims at higher-order thinking and problem solving.
- Diverse learners are given equal opportunity.
- Learners are socialized into the discourse and practices of the academic disciplines.
- The relationship between learning in and out of school is authentic.
- Important dispositions and habits of mind are fostered.
- Democratic practices are enacted within a caring community.

(b) Cognitive and Constructivist Learning Theories

- Intellectual abilities are socially and culturally developed.
- Learners construct knowledge and understandings within a social context.
- New learning is shaped by prior knowledge and cultural perspectives.
- Intelligent thought involves "metacognition" or self-monitoring of learning and thinking.
- Deep understanding is principled and supports transfer.
- Cognitive performance depends on dispositions and personal identity.

(c) Classroom Assessment

- Challenging tasks elicit higher-order thinking.
- Learning processes and learning outcomes are addressed.
- Assessment as an on-going process, integrated with instruction.
- Assessment is used formatively to support student learning.
- Students clearly understand expectations.
- Students actively evaluate their own work.
- Both student learning and teaching are evaluated.

The key tenets of Shepard's (2000) "emergent constructivist" paradigm are distinctly different from those espoused by behaviorist theories, specifically the latter's focus on hereditarian theories of intelligence and ability. Uppermost in Shepard's (2000, p. 8) "reformed" framework is the belief that "all students can learn." The features of the "reformed" curriculum vision are closely aligned to the essential characteristics of constructivist theories, including challenging subject matter aimed at higher-order thinking and problem solving. Alluding to the importance of the context for learning, she envisages authenticity in the relationship between learning in and out of school (Shepard, 2000). In the "reformed" curriculum vision, "instructional conversations" (Shepard, 2001, p. 1078) between teachers and students serve the purpose of sharing information, developing common meanings, and socializing students to the nature of reasoning and thinking within a disciplinary field. The next feature, 'authenticity in the relationship between learning in and out of school,' has parallels with 'authentic academic achievement' (Newmann & Associates, 1996). For students to realize authentic academic achievement, *Disciplined Inquiry* (Newmann & Associates, 1996, p. 25) is an important criterion because accomplishments are developed from prior knowledge that has been accumulated within a disciplinary field of knowledge. In terms of teaching, the enactment of this curriculum framework in the classroom is based on cognitive and constructivist learning theories. Specifically, Shepard (2000) envisions six defining tenets of teaching – all mirroring closely the characteristics of teaching and learning within cognitive and sociocultural constructivist theories [Figure 2.1(b)].

The envisaged assessment practices in this "emergent constructivist" paradigm point to the importance of 'classroom assessment' [Figure 2.1(c)]. A significant feature is for assessment to be a continuing process that is closely integrated into instruction (Shepard, 2000, 2001). A

second important feature is *formative assessment*, which, in Shepard's model, is part of supporting student learning. As discussed earlier, constructivist learning theories conceive of students as building on and extending prior knowledge over time. Due to this developmental emphasis, formative assessment, which is typically used to help students to close their learning gaps, is an important teaching tool.

Situating constructivist theories in the study context

This dissertation suggests that the intent of TSLN-TLLM is consistent with the essential features and characteristics of constructivist learning theories. To help teachers understand the philosophy and spirit of TSLN-TLLM, the Singapore MOE website writes that TLLM is intended to improve the "quality of interaction between teachers and students (MOE [Bluesky], 2005). This is achieved by revisiting the *what*, *why*, and *how* of teaching; it is a timely reminder that education is to prepare students for the "test of life" and not a "life of tests" (MOE [Bluesky], 2005).

Table 2.1 provides an overview of the TLLM vision that shapes the *what*, *why* and *how* of teaching and aligns it to corresponding features of Shepard's (2000, 2001) "emergent constructivist" paradigm. Overall, MOE's TLLM tenets have striking similarities with constructivist learning theories. For instance, as presented in the table below, under *why we teach,* TLLM focuses on the learner, as does Shepard's curriculum framework which believes that "all children can learn." Aspects in Shepard's "reformed" curriculum, such as *challenging subject matter aimed at higher-order thinking and problem solving*, and *the relationship between learning in and out of school is authentic*, are similar to the TSLN pillars of national education and thinking skills.

Table 2.1
Comparison of TLLM with the "Emergent constructivist paradigm" (Shepard, 2000)

TLLM Vision		**Shepard's "Emergent Constructivist" Paradigm**[a]
More …	**Less …**	
Remember why we teach		*Curriculum*
For the learner	To rush through the syllabus	All students can learn.
To excite passion	Out of fear of failure	
For understanding	To dispense information only	Challenging subject matter aims at higher-order thinking and problem solving.
For the test of life	For a life of tests	The relationship between learning in and out of school is authentic.
Reflect on what we teach		
The whole child	The subject	All students can learn.
Values-centric	Grades-centric	Important dispositions and habits of mind are fostered.
A process	A product	Learning processes and learning outcomes are addressed.
Searching questions	Textbook answers	Challenging subject matter aims at higher-order thinking and problem solving.
Reconsider how we teach		
Engaged learning	Drill and practice	
Differentiated teaching	'One-size-fits-all' instruction	All students can learn.
Guiding, facilitating, modeling	Telling	Learners construct knowledge and understandings within a social context.
Formative and qualitative assessing	Summative and quantitative testing	Assessment as an on-going process, integrated with instruction. Assessment is used formatively to support student learning.
Promoting a spirit of innovation and enterprise	Insisting on set formulae, standard answers	Challenging subject matter aims at higher-order thinking and problem solving.

[a] Adapted from the "emergent constructivist" paradigm (Shepard, 2000)

In terms of teaching, Shepard's pedagogical vision in the "emergent constructivist" framework calls for learners to construct knowledge and understanding within a social context, develop deep understanding, and to engage in metacognition. These features resonate with the tenets of *how we teach* in TLLM (e.g., differentiated learning, guiding, facilitating, modeling).

Finally, the intent for assessment in Shepard's "emergent constructivist" vision calls for *higher-order thinking*, *emphasizes the process as well as products (or outcomes) of learning, is*

more tightly integrated into teaching, *uses assessment evidence and data formatively to further student learning*, and *provides opportunities to involve students in actively evaluating their own work* (Shepard, 2001). These characteristics echo the TSLN philosophy of assessing for understanding, focusing on the process of learning, and emphasizing formative and qualitative assessing (MOE [Bluesky], 2005).

Summary

Using constructivist learning theories and their implications for assessment as a theoretical perspective, I will interpret and analyze the concepts of *formative assessment* (Black & Wiliam, 1998a) and *authentic intellectual work* (Newmann & Associates, 1996; Newmann, Marks, & Gamoran, 1996), to examine how these interact and play out in Singapore teachers' classroom assessment practices. Specifically, this dissertation uses the criteria in Newmann and Associates' (1996) authentic intellectual achievement as proxy indicators of the type of 21st century learning envisaged in the TSLN-TLLM visions. It also explores and formative assessment as a means of realizing the vision of helping students *learn more* as envisioned in TLLM.

Conceptual orientation: Classroom assessment, Authentic Assessment, Formative Assessment

The conceptual orientation section examines *authentic intellectual work* (AIW) (Newmann & Associates, 1996; Newmann, King, & Carmichael, 2007) and *formative assessment* (Black & Wiliam, 1998a). In this dissertation, both concepts are examined in relation to assessment in a constructivist paradigm, and are discussed under the umbrella of classroom assessment. This section begins with a discussion of classroom assessment, and its

history and use, focusing on its prevalence, decline, and current ascendency, as well as its relation to authentic intellectual work and formative assessment.

Next, I delineate and discuss the contested and contentious concept of *authentic assessment*, and present the conceptual and empirical work that examines the quality of teacher assessment, as well as the relationship between it and student learning. This dissertation uses the quality of student work captured under the AIW concept as a proxy for student learning.

In the third section, I discuss "formative assessment," as conceptualized by Black and Wiliam (1998a). These two scholars popularized this aspect of classroom assessment with their seminal review of the literature on the relationship between teacher assessment and student achievement. The conceptual and empirical work on formative assessment is also presented. Finally, this chapter concludes by integrating authentic intellectual work, formative assessment and constructivist learning theories, identifying the gaps in the research, and contextualizing and explaining the focus of the research for the dissertation.

Classroom assessment

Definition

This dissertation is concerned with examining assessment practices that occur in the classroom, and that are part of the activities integral to teaching and learning. This dissertation uses three terms – classroom assessment, formative assessment, and authentic assessment – which are differentiated in this section. For this dissertation, I use *classroom assessment* to encompass teacher assessment, an overarching term that encapsulates different purposes (i.e., summative and formative), theories of learning, and assessment types (e.g., modes, and formats).

This dissertation defines *authentic assessment* as assessments used by teachers in the classroom that requires students to construct knowledge through disciplined inquiry, and apply

what is learned to a context that has value beyond school. To this end, this dissertation uses authentic assessment as a particular type of classroom assessment developed and designed by teachers to ascertain students' ability to build on prior knowledge and apply this information to new contexts, typically seen by scholars as "real world" situations. This definition is drawn from the AIW criteria (Newmann & Associates, 1996), which requires that assessment tasks be developed based on three specific criteria: *Construction of Knowledge*, *Disciplined Inquiry*, and *Value Beyond School*. These characteristics of AIW resonate with the types of skills with which Singapore students should be equipped with when they leave formal education, according to the TSLN vision.

The third concept, *formative assessment* is broadly conceived as "all those activities undertaken by teachers and/or by their students, which provide information to be used as feedback to modify the teaching and learning activities" (Black & Wiliam, 1998a, p. 7). This definition implies that formative assessment is closely integrated into teaching, and serves to provide information to *both* teachers and students, not just to teachers. This conception of formative assessment resonates strongly with constructivist learning theories in that students incorporate new information based on the assessment to improve their learning, are active in the learning process, and hence are seen as taking charge of their learning after receiving feedback from the assessment process.

History

Classroom assessment has had a somewhat turbulent history, and after falling out of favor for several decades, its value in teaching and learning appears to be on the rise once more. Since the 1960s, following the move toward increased accountability demands made on public schools in the USA and UK, standardized national assessments across states and school districts became

prevalent (Stiggins, 2002). Much of the shift away from teacher assessment was due to the wide variability in terms of the quality of teacher-designed tasks (Crooks, 1988) and grading practices (Cizek, 1997), the lack of teacher training and expertise in assessment (Leighton, Gokiert, Cor, & Heffernan, 2010; Stiggins, 1991; Tanner, 2001; Wolf, et al., 1991), and the inability to implement assessment-related activities (Plake & Impara, 1997). Furthermore, teachers were not cognizant of the incongruity between their intended and enacted assessment practices (Bol & Strage, 1996). To this end, it may be argued that teachers have been unknowing contributors to the ascendency of standardized testing (Wiggins, 1989).

In recent years, there appears to be a policy shifts that indicate an acknowledgment of the value of teacher assessment (Leung & Rea-Dickins, 2007). One reason for this growing recognition is the recent research that identifies the misalignment between new learning theories, and the limitations of formal external assessments (Eisner, 1991; James, 2006). Another reason is the awareness that teachers devote a significant amount of time in the classroom to assessing student learning (T. L. Good & Brophy, 2008; Pellegrino & Goldman, 2008; Stiggins, 1992; Suah & Ong, 2012). This perception may also be due to teachers preferring to rely on their own observations in determining what their students know and are able to do (Camp, 1992). Ideally, because classroom assessments are conducted during instruction, they provide the most meaningful information to teachers in order to identify what they have taught well and what they need to revisit (Guskey, 2003). This is important because teachers spend as much as 50 percent of classroom time assessing students, formally or informally (T. L. Good & Brophy, 2008), and because teacher assessments make up a significantly large proportion of students' experiences in school (Brookhart, 2001; Mertler, 1999).

Empirical work

A review of the recent peer-reviewed literature (and other sources) on teachers' assessment practices from 1998 – 2011 shows several different patterns of classroom assessment practices. The empirical work generally takes the form of observation studies and shows five patterns in teacher assessment practices: (1) by grade level (Bol, et al., 1998; Koh et al., 2005; Koh & Luke, 2009; McNair, Bhargava, Adams, Edgerton, & Kypros, 2003; Mertler, 1999; Zhang & Burry-Stock, 2003b); (2) by content area (Bol, et al., 1998; Koh, et al., 2005; Koh & Luke, 2009; McMillan, 2001; Zhang & Burry-Stock, 2003b); (3) by assessment types (Koh, et al., 2005; Koh & Luke, 2009; McMillan, Myran, & Workman, 2002; Mertler, 1999; Ohlsen, 2007); (4) by types of skills assessed (McMillan, 2001; McMillan, et al., 2002); and (5) by teacher experience (Bol, et al., 1998).

In terms of grade levels, there were differences between the types of assessment used by elementary and high school teachers. Elementary teachers use alternative assessment more frequently than high school teachers, while middle and high school teachers use traditional assessment more frequently than elementary teachers (Mertler, 1999). Elementary teachers were also reported to use informal observations, portfolios and questioning more frequently than middle and high school teachers. The research indicated that teachers teaching higher grades tended to use more objective paper-and-pen tests because of concerns about assessment quality and of the need to prepare students for state testing (McNair, et al., 2003).

Some studies report that teachers' classroom assessment practices vary by content area (Koh, et al., 2005; McMillan, et al., 2002; Zhang & Burry-Stock, 2003a). In a study involving the grading and assessment decisions of 900 Grades 3-5 teachers in Virginia, McMillan et al. (2002) concluded that mathematics teachers used performance assessments and projects less

regularly than language arts teachers. Generally, teachers overseeing academic subjects are more involved in specific assessment activities than those teaching non-academic subjects since the former subjects are mandated for state testing. Thus, mathematics teachers have a higher tendency to use major examinations and objective assessments than language arts teachers (Zhang & Burry-Stock, 2003a). They also assign more homework (Koh, et al., 2005). But interestingly, in Singapore, Koh et al. (2005) found that social studies teachers assign more homework than their counterparts teaching English or science. Using a methodology that involved the collection of samples of teacher assignments from 36 Singapore schools, Koh et al. (2005) found that, regardless of level (either Primary 5 or Secondary 3), teachers who teach mathematics and social studies assign more homework than teachers who teach English, science, biology or physics.

Patterns of teacher assessments also varied according to the nature of the test. Traditional objective assessments are used more frequently than alternative assessments (McMillan, et al., 2002; Mertler, 1999). However, when the analyses were disaggregated based on teachers' experience, the most experienced teachers (those with more than 20 years of teaching experience used) alternative assessments more frequently than the least experienced teachers (those with less for than six years of experience) (Bol, et al., 1998). One pattern reported that mathematics teachers use traditional assessments less frequently than do their colleagues teaching other disciplines (Bol, et al., 1998). One possible reason for this is that the emphasis on mathematical problem-solving precludes the use of traditional types of assessment (Bol, et al., 1998).

Teacher classroom assessment is divided between lower order (e.g. recall of knowledge and concepts) and higher-order skills (e.g. focus on application and reasoning). For instance, McMillan et al. (2002) examined elementary teachers' assessment of cognitive skills and

reported four types of skills which were commonly assessed: recall, understanding, reasoning and application. Both mathematics and language arts teachers value the assessment of application skills over recall, indicating that they value higher-order skills over recall and knowledge-reproduction. More details on teachers' assessment of higher-order skills will be discussed under the review of the empirical work related to AIW.

Teacher assessment practices are based on the a national curriculum, especially if this curriculum provides explicit details on the new direction of assessment. To examine Grade 7-10 teachers' responses to a new mathematics curricular directive in Ontario, Canada, Surrtamm, Koch, and Arden (2010) administered a self-report questionnaire to survey 1096 mathematics teachers' classroom assessment practices and reported that teachers were using a variety of different assessments congruent with the mathematics reform practices. Specifically, the participating teachers were using a variety of assessment forms (e.g., pencil-and-paper tests, quizzes), adopting seamless instruction and assessment practices, focusing on the complexity of the task (e.g., mathematical thinking), and emphasizing formative assessment.

Summary

The empirical work reports a diverse array of classroom assessment practices. The researchers used questionnaire surveys and reported findings which were statistically significant because large samples of teachers had contributed responses. What is missing from this set of studies published between 1998 and 2011 is the use of qualitative research to examine teachers' rationales for and experiences of using such classroom assessment practices. In addition, within the peer-reviewed journals, only a few studies (e.g., Suurtamm, et al., 2010) examine the nature and pattern of teachers' classroom assessment within the educational change process. Therefore,

this dissertation study contributes to the extant epistemology by examining classroom assessment practices within Singapore's educational change process.

Conceptualizing Authentic Intellectual Work

Definition.

The word, 'authentic' has Greek and Old French roots. In Greek, 'authentikos' refers to "original, genuine, principal" while in Old French, 'authentique,' means "canonical" (Online Etymology Dictionary, 2013). Dictionaries generally define 'authentic' as something that is "genuine" or "as good as the original" (Collins Cobuild Advanced Learner's English Dictionary, 2005), or "conforming to the original so as to reproduce essential features" (Merriam-Webster, 2011). In 'authentic assessment,' Newmann and Associates (1996, p. 22) view 'authentic' as "something that is real, genuine, or true rather than artificial, fake, or misleading." However, defining 'authentic' by comparing it to its antonym (e.g. Newmann and Associates' use of "artificial," and "fake") is problematic. In fact, 'authentic' education is "unspecified" (Gulikers, et al., 2004, p. 67) and is among a number of "ill-defined concepts and terms" in education and educational research (Palm, 2008, p. 1) today.

The call to use authentic assessment is in response to the limitations of standardized, objective tests. Some of the limitations are the overemphasis on the assessment of bite-sized, disconnected and disparate pieces of facts; the focus on rote, recall, replication, and regurgitation of knowledge (Cole, 1990; Shepard & Kirst, 1991); the privileging of just one indicator of achievement (Archbald & Newmann, 1988); and the reliance on assessment formats (such as multiple choice and short response items) that provide limited opportunities for students to demonstrate communication skills (Newmann, Lopez, & Bryk, 1998).

The shortcomings of objective standardized assessments means that students are not assessed on higher-order thinking and in-depth understanding (Newmann, Lopez, et al., 1998), that assessment provides an incomplete picture of student performance (Newmann, Lopez, et al., 1998), and that students do not appreciate or conceive of how assessment, learning and school work may produce important and meaningful outcomes (Newmann & Associates, 1996). In the 21st century, knowledge is seen as contested and complex, and this conflicts with the use of multiple-choice and short answer responses which give students the misconception that there is always one right answer (Wiggins, 1989).

From a teaching and learning perspective, there is a backwash effect when standardized tests drive the curriculum (Camp, 1992; Gulikers, et al., 2004; D. A. Watkins & Biggs, 2001; Wiggins, 1989). Because standardized tests focus largely on literacy and numeracy, for example, subjects such as social studies, physical education, and art, considered as "frills" in the curriculum are neglected (Shepard & Kirst, 1991, p. 21). Multiple-choice items which are both easy to score and cost effective to produce are overused and further influence the curriculum. Teachers are able to mimic this test format easily in their daily assignments (Shepard & Kirst, 1991), and consequently merely drill students on these test strategies.

Despite the limitations of standardized tests, authentic intellectual achievement supporters do not insist that all school assessments imitate work or activities outside of school (Newmann, et al., 1996); rather they urge educators to consider the different aspects of authentic intellectual quality in education (Bryk, et al., 2000), and acknowledge that there are benefits in combining assessments that require students to construct as well as recall knowledge (Christenson, 1991).

History and theoretical conceptualization

Although the focus on authentic assessment is a recent phenomenon, authenticity in education dates back to Aristotle (Splitter, 2009). While Grant Wiggins has been credited as being the most influential contemporary advocate of 'authentic assessment' (Terwilliger, 1997), Archbald and Newmann (1988) are said to have predated Wiggins when, in 1988, they published a vision of 'authentic achievement' (see Newmann, Brandt, & Wiggins, 1998, for the exchange of views on semantics and psychometrics relating to authentic assessment). There is a semantic difference between 'assessment of authentic achievement' and 'authentic assessment of achievement' (Cumming & Maxwell, 1999). The use of 'assessment of authentic achievement' places a premium on the nature of the achievement, while 'authentic assessment of achievement' indicates an emphasis on the approach to assessment, and this may not examine the nature of the achievement itself (Cumming & Maxwell, 1999). This dissertation uses both terms in the following ways:

- Assessment of authentic achievement – This will be used synonymously with "authentic achievement" (Newmann & Associates, 1996) to refer to the types of higher-order thinking skills required for the 21st century. This dissertation conceives of this term as the outcomes that the assessments seek to assess. Additionally, this dissertation uses *assessment of authentic achievement* synonymously with *authentic academic achievement* (Archbald & Newmann, 1988).
- Authentic assessment of achievement – This will be applied to the types and nature of teacher assessment that this dissertation seeks to examine. Specially, this dissertation conceives of these types of assessment being designed and examined with respect to the AIW criteria and standards (Newmann, et al., 2007). Further details of the standards and criteria are provided below. This dissertation uses *authentic assessment of achievement* synonymously with *authentic intellectual work* and *authentic assessment*.

Authentic assessment can take many modes, including essays, reports, performances and portfolios (Wiggins, 1989). However, it is the purpose and not the mode of the assessment that is important. Authentic assessments are premised on the idea that "the test is central to

instruction" (Wiggins, 1989, p. 704) and "assessment tasks … are real instances of extended criteria performances, rather than proxies or estimators of real learning goals" (Shepard & Kirst, 1991, p. 21). In this way, authentic assessment is an integral part of instruction, rather than the driver of instruction (Wiggins, 1989). An authentic assessment has four criteria (Wiggins, 1989):

- The task is designed to be representative of its performance in the field and the activities are based on ill-structured challenges that do not have clear solutions. This criterion is intended to prepare learners for the complexities of life outside school, making them aware that there is no single correct response, unlike the standard multiple choice response which has a single right answer.
- The criteria used must be well-defined within the area of expertise, and must closely parallel what is 'essential' in the field.
- Student self-assessment musts play a role such that students have the opportunity to review, revise and redirect their work and learning.
- Learners are required to present their work, publically, and orally. This last criterion serves to motivate learners and to signal to them that their work is sufficiently significant to warrant an audience.

Drawing on the early conceptions of authentic assessment, this term may be defined according to the following categories: (1) by 'what it is not' (Wiggins, 1990); (2) by comparing it to traditional assessments (Tanner, 2001); (3) by using foci and perspectives (Palm, 2008; Wiggins, 1989); and (4) as a philosophy about classroom assessment (Elliott, 1991). Within this dissertation, the broad working definition of 'authentic assessment' is taken as

> intellectual accomplishments that are worthwhile, significant, and meaningful, such as those undertaken by successful adults: scientists, musicians, business entrepreneurs, politicians, crafts people, attorneys, novelists, physicians, designers, and so on (Newmann & Associates, 1996, pp. 23-24).

Fred Newmann developed and further delineated the ideals of authentic achievement in his work with different scholars. From the broad vision articulated with Doug Archbald in 1988, he went on—with colleagues from the Centre on Organization and Restructuring of Schools—to refine the three components of authentic achievement, created standards for applying these

criteria to authentic achievement, and developed an extensive methodological approach to examine instruction, assessment task, and student work based on the criteria.

'Authentic achievement' according to Newmann and Associates (1996) comprises three **components**: instruction, assessment tasks, and student performance. The first two components together make up authentic pedagogy. The use of 'authentic achievement' is also an update of 'authentic academic achievement' which was first used by Archbald and Newmann (1988). The three components are defined by three **criteria**: *Construction of Knowledge*, *Disciplined Inquiry*, and *Value Beyond School* (Newmann & Associates, 1996). Putting these together, we have the following picture.

Authentic Achievement **Components**
= Authentic Pedagogy + Authentic Student Performance
(Authentic Instruction + Authentic Assessment) + Authentic Student Performance.

Authentic Achievement **Criteria**: (Construction of Knowledge, Disciplined Inquiry, Value Beyond School)

Briefly, *Construction of Knowledge* is about the application and transfer of knowledge, rather than its reproduction. Similar to the skills articulated in the 21st century skills framework, construction of knowledge focuses on higher-order thinking skills, where students demonstrate their ability to synthesize, analyze, and evaluate knowledge and facts, and in so doing, they develop deep understanding of subject-matter (Newmann, Secada, & Wehlage, 1995), allowing them to "arrive at conclusions that produce new meanings and understandings" of the material learned (Newmann, et al., 1996, p. 285).

The second criterion, *Disciplinary Inquiry*, emphasizes the role of subject knowledge similar to the view emanating from the 'scholar academic' (Kliebard, 2004; Schiro, 2008) notion of the curriculum, in that it conceives of learning as adding to the extant canon of knowledge

(Newmann & Associates, 1996). The goal of *Disciplinary Inquiry* is for learning to be more than the superficial acquisition of discrete bits of knowledge, and for it to transcend knowledge as facts, concepts, and theories toward developing deep understanding (Bryk, et al., 2000; Newmann & Associates, 1996; Newmann, et al., 1995).

The third criterion, *Value Beyond School*, refers to assessments that require students to be engaged in activities in which adults in the real world participate. These assessments are contextualized so that students can apply and transfer what they learn to a genuine setting mirroring adult activities. This criterion was developed in response to the critique of standardized tests in that there is little opportunity for students to apply what is learned (Newmann & Associates, 1996). Comparatively, authentic assessments have "aesthetic, utilitarian, or personal value" (Newmann & Associates, 1996, p. 26). Based on these criteria, it is evident that the definitive goal of authentic intellectual achievement is

> to cultivate the kind of higher-order thinking and problem-solving capacities useful to both individuals and to the society. The mastery gained in schools is likely to transfer more readily to life beyond school (Newmann & Archbald, 1992, p. 75)

There are similarities between Wiggins' (1989), and Archbald and Newmann's (1988) conceptions of authentic assessment in that both perceive the need for collaboration (a skill valuable in the world beyond school), for deep understanding, and for a grasp of disciplinary knowledge. There are also differences in that Wiggins (1989) includes formative assessment as part of the assessment process. Together, these conceptions of authentic assessment echo the characteristics of assessments within constructivist learning theories. The similarities are that assessment should be contextualized and socially constructed, that assessments should move beyond responding to items within a limited time frame to participating in activities and tasks that take place over time, and that assessments provide the opportunity for peer collaboration.

Such competencies and interactions are what learners will encounter outside of school, and therefore must be assessed. In a parody commenting on educational reform, Sir Ken Robinson (2010) remarks that contrary to the value of collaboration in the outside world, working with peers is often seen as cheating in schools. This statement made over twenty years after Wiggins (1989), and Archbald and Newmann (1988) delineated their vision for authentic assessment is a stark indicator of the divide that still persists between the skills, knowledge and competencies measured in schools and that which is is required and valued by the world outside of school.

In order to examine whether teacher assessments met the authentic assessment criteria, Newmann and Associates (1996) developed 14 standards to evaluate teaching and assessments based on the three criteria. There are 7 standards for authentic assessment tasks designed or used by teachers, 4 for authentic instruction, and 3 for authentic student performance. These standards for evaluating authentic achievement are shown in Table 2.2.

Table 2.2
Standards for authentic achievement, authentic pedagogy and authentic student performance (Newmann & Associates, 1996)

Authentic achievement criteria [a]	Authentic pedagogy		Authentic student performance
	Authentic assessment tasks	Authentic instruction	
Construction of Knowledge	Organization of information Consideration of alternatives	Higher-order thinking	Analysis
Disciplined Inquiry	Content Process Elaborated written communication	Deep knowledge Substantive conversation	Disciplinary concepts Elaborated written communications
Value Beyond School	Problem connected to the world beyond the classroom Audience beyond the school	Connections to the world beyond the classroom	

[a] Adapted from Newmann and Associates (1996, p. 46). While student performance would ideally be evaluated for *Value Beyond School*, Newmann and Associates did not develop a standard for this because their research was unable to include interviews with students to ascertain how they conceived of the assessment tasks.

Over the years, Fred Newmann has refined and revised the 'authentic achievement' nomenclature and standards. A compilation of the terms and standards from 1988 to 2007 is presented in Table 2.3. While the terms and their descriptions have changed over the years, there is no documentation in the literature for the reasons behind these modifications. Beginning in 1988, Archbald and Newmann (1988) used 'authentic academic achievement' as a means of looking at assessment. Subsequently, Newmann and Wehlage (1993) introduced the term, 'authentic instruction' as a way of thinking about teaching that engages students and enhances the way they think. This led to the combination of 'authentic instruction and assessment' (Newmann, et al., 1995). By 1996, authentic achievement was introduced to encapsulate authentic pedagogy (authentic assessment task and authentic instruction), and authentic student performance. This indicates that by the fourth iteration, Newmann and colleagues had deepened their conceptualization of 'authentic assessment' beyond the broad vision to identify specific terminology for teacher assignments (authentic assessment tasks), pedagogy (authentic instruction), and student work (authentic student performance). In 2000, Authentic Intellectual Work was introduced in a report to the Iowa Department of Education, and defined as intellectual work that "involves original application of knowledge and skills, rather than just routine use of facts and procedures," as well as the "careful study of the details of a particular problem and results in a product or presentation that has meaning beyond success in school" (Newmann, et al., 2007, p. 3).

Other refinements are germane to the standards for assessing Authentic Intellectual Work. Some of the revisions made include introducing 'higher-order thinking' in place of 'integration of knowledge' for Construction of Knowledge between 1988 and 1993 and the inclusion and subsequent removal of 'consideration of alternatives' under *Construction of Knowledge*. This

dissertation will apply these standards from the latest iteration to rate the quality of Singapore teachers' assessment tasks.

Table 2.3
Developments in authentic achievement

		Criteria		
	Nomenclature	**Construction of Knowledge**	**Disciplined Inquiry**	**Value beyond School**
Archbald & Newmann (1988)	Authentic academic achievement	Integration of knowledge[a]	Disciplined inquiry • Prior substantive and procedural knowledge • In-depth understanding • Produce knowledge, assemble and interpret information	Value beyond evaluation • Production of discourse, things, performances • Flexible use of time • Collaboration
Newmann & Wehlage (1993)	Authentic Instruction	Higher-order thinking, Social support for student achievement	Depth of knowledge Substantive conversation	Connectedness to the world
Newmann, Secada, & Wehlage (1995)	Authentic instruction and assessment (also authentic achievement)	Higher-order thinking	Deep knowledge Substantive conversation	Connections to the world beyond the classroom
Newmann & Associates (1996)	Authentic achievement (for assessment tasks)	Organization of information Consideration of alternatives	Content Process Elaborated written communication	Problem connected to the world beyond the classroom Audience beyond the classroom
	Authentic achievement (for instruction)	Higher-order thinking	Deep knowledge Substantive conversation	Connections to the world beyond the classroom
	Authentic student performance	Analysis	Disciplinary concepts	Elaborated written communication
Newmann, Marks, & Gamoran (1996)	Authentic pedagogy (authentic assessment tasks)	Organization of information Consideration of alternatives	Content Process Elaborated written communication	Problem Audience
	Authentic pedagogy (Authentic instruction)	Higher-order thinking	Deep knowledge Substantive conversation	Connections to the world beyond the classroom
	Authentic student academic performance	Analysis	Disciplinary concepts Elaborated written communication	
Newmann, Lopez & Bryk (1998)	Authentic intellectual work (AIW)	Apply or extend prior knowledge	Prior knowledge base In-depth understanding Elaborated communication	Connections to students' lives
Newmann, King & Carmichael (2007)	Authentic instruction and Assessment Authentic intellectual work (AIW)	Authentic assignments: Construction of knowledge	Authentic assignments: Elaborated written communication	Authentic assignments: Connections to student lives
		Authentic instruction: Higher-order thinking	Authentic instruction: Deep Knowledge substantive conversation	Authentic instruction: Connections to the world beyond the classroom
		Student work: Analysis	Student work: Disciplinary concepts Elaborated written communication	

[a] The criterion, 'Integration of knowledge' was used in place of 'Construction of knowledge' in Archbald and Newmann (1988).

Newmann and Associates (1996) maintain that authentic assessment must embrace all three criteria. Assessment and learning tasks that are lacking in one or more criteria are not authentic. Thus, while authentic assessment is frequently associated with both performance and alternative assessment (Herman, et al., 1992), or just performance (Palm, 2008) or alternative (A. Hargreaves, Earl, & Schmidt, 2002) assessment, the three terms are not synonymous. While alternative assessments may differ from the traditional pen-and-paper assessments, the tasks may not be situated in a 'real world' context (S. Smith, Layng, & Jones, 1995). Tasks that meet the requirement of a 'real world' context typically have four criteria, namely, they have a personal frame of reference, lack a clear solution for solving the problem or issue, motivate the learner to explore options, and address a target audience (Rezulli, Gentry, & Reis, 2004). Likewise, performance assessments may require a performance but this may also not fit the criteria of authentic achievement.

Noting their stringent criteria, Newmann and Associates (1996) are cognizant of the constraints in classrooms, and clarify that they do not expect teaching and assessment activities to meet the three criteria all of the time (Newmann, et al., 1995). After all, repetitive and routine practices and memory skills are required to build foundational knowledge and skills that would enable the use of such authentic tasks (Newmann, et al., 1995). Furthermore, research on the science of learning recognizes that facts are important for thinking and problem solving (Bransford, et al., 2000). In fact, experts' abilities to critically analyze and problem-solve depend on a deep knowledge of the subject matter (Bransford, et al., 2000). International benchmarking studies like TIMSS include *knowing* as one of the cognitive domains because factual knowledge enables students to engage in more complex cognitive tasks (Mullis, Martin, Ruddock, O'Sullivan, & Preuschoff, 2009). Finally, scholars (e.g., Christenson, 1991; Tanner,

2001) do not advocate completely discarding traditional assessment in favor of authentic assessment but champion a blended approach for teachers.

Conceptually, the authentic assessment framework has continued to be influential over the years, with scholars applying and adapting the framework. For example, Callison and Lamb (2004, p. 34) formulated "signs of authentic learning" to be found where there is student-centered learning, use of multiple resources accessible outside the school, use of process, product, and performance assessment, as well as increased collaboration during teaching and learning. They also proposed that authentic assessments contain "authentic context," "authentic questions" and "authentic tasks," and the use of "authentic communication and audiences" to support the authentic assessments.

Other theoretical lenses have been used to extend the authentic achievement framework. Hayes, Mills, Christie, and Lingard (2006) drew on work on school reform, critical theory, sociolinguistics, feminism, sociology of education, and critical pedagogy in order to create a nomenclature more relevant for their own use. More specifically, they preferred to use "productive outcomes" instead of 'authentic intellectual achievement' to examine both intellectual and social outcomes of education. Similarly, Gulikers et al. (2004) drew on literature on authentic assessment, authenticity, and assessment in general, as well as student perceptions of (authentic) assessment to develop a five-dimensional framework for authentic assessment based on the (a) assessment task, (b) physical context, (c) social context, (d) assessment or form, and (e) assessment criteria.

Geographically, the influence of authentic intellectual work was extended to include research in the Consortium on Chicago School Research as part of the Chicago Annenberg Challenge in the late 1990s and early 2000s. Other related studies in the USA which apply

Newmann and Associates' conceptualization of authentic assessment include the school restructuring studies using data from the National Educational Longitudinal Study data (see V. E. Lee & Smith, 1994; V. E. Lee, Smith, & Croninger, 1995; V. E. Lee, Smith, & Croninger, 1997), research on the achievement of youth with disabilities (see Braden, Schroeder, & Buckley, 2001; Hanley-Maxwell, Phelps, Braden, & Warren, 1999; King, Schroeder, & Chawszczewski, 2001), and also studies in Minnesota (see Avery, 1999; Avery, Jouneski, & Odendahl, 2001). In Queensland, Australia, Bob Lingard and colleagues adopted and adapted Newmann and Associates' framework (see Hayes, et al., 2006; Lingard, Mills, & Hayes, 2006) and applied this to examine the quality of instruction and assessment in schools. Allan Luke, who was a member of the Queensland team, introduced Newmann and Associates' methodology to a large-scale study of reform in Singapore (see Koh & Luke, 2009; Luke, Freebody, Shun, & Gopinathan, 2005; Luke & Hogan, 2006). More recently, scholars in New Zealand have started examining Newmann and Associates' notion of authenticity within the country's 2007 technology education curriculum (Snape & Fox-Turnbull, 2011). A chronological summary of this development is presented in Table 2.4.

Table 2.4
Research studies applying and adapting Newmann and Associates' (1996) framework

Year of research	Geographical location	Research team	Name of study
1990-1994	USA	Newmann & Associates	Center for Organization and Restructuring of Schools School Restructuring Study (CORS-SRS)
1996-1999	Chicago, USA	Bryk, Nagaoka, & Newmann (2000) Newmann, Lopez, & Bryk (1998)	Chicago Annenberg Research Project (CARP)
1999-2003	USA	King, Schroeder, & Chawszczweski (2001)	Research Institute on Secondary Education Reform for Youth with Disabilities (RISER)
1998-2000	Queensland, Australia	Lingard, Mills, & Hayes (2006)	Queensland School Reform Longitudinal Study (QSRLS)
2004-2007	Sydney, Australia	Ladwig, Gore, Amosa, & Griffiths (2007)	Systemic Implications of Pedagogy and Achievement in New South Wales Public Schools Study (SIPA)
2004-2005	Singapore	Luke, Freebody, Shun, & Gopinathan (2005) Koh and Luke (2009)	Core Panel Study (Panel 5)

Debates and challenges.

Despite the careful attempts Newmann and Associates took to define and delineate the AIW concept, there were a number of substantive, semantic, philosophical, psychometrical, and practical issues that proved problematic. In particular, the choice of the adjective, 'authentic' was contentious, and this resulted in many philosophical and educational debates over its meaning.

Semantic issue. The meaning of 'authentic' is hotly disputed. Terwilliger (1997) insists that the word 'authentic' almost has "mystical power" (p. 24). He expresses discomfort with the use of the word because its antonym is "inauthentic," and he ponders the implications for instruction, performance, and outcomes that might be labeled as such. Gresham's (1991) misgivings lie with another antonym of 'authentic,' "counterfeit." A second semantic issue arises from the synonyms of 'authentic,' which include "real," "realistic," and "real life"

(Terwilliger, 1997). This makes the definition "circular" (Terwilliger, 1997, p. 25). Because the word is value-laden, 'authentic' should be replaced by a less contentious and more "neutral" term (Terwilliger, 1998, p. 23). Finally, Terwilliger takes issue with authentic assessment advocates' advancement of higher-order thinking skills because this suggests that they are contemptuous of *basic skills*, which, in his opinion, also have their place in education. In response, Newmann et al. (1995) clarify that they do not dismiss other aspects of student learning but reinforce their view that the criteria for authentic assessment emphasize depth more than breadth in the teaching and learning of the curriculum, and assert that this in-depth exploration enhances students' learning.

Philosophical issue. Within a postmodern context, the philosophical meaning of 'authentic' as well as 'authentic assessment' has been debated. For instance, can 'authentic' mean something that is genuine, and that is not contradicted by evidence (see for example, A. Hargreaves, et al., 2002)? Schools are diverse places, and students in schools all hold multiple perspectives (A. Hargreaves, et al., 2002). As a result, 'authentic' is subjective and relative to context. This raises the question of what is true authentic work for each and every student because "what accounts as authentic for one person may be far from authentic for another" (Splitter, 2009, pp. 137-138). The question then is whether 'authentic' assessments and learning can be developed for each and every student. In addition, since teachers design and develop assessment tasks, these tasks reflect the teachers' understanding of authenticity, which differs from the understanding of other teachers, and of students (Splitter, 2009).

According to Resnick (1987a), the authenticity of education in schools is questionable in four aspects because schools are an artificial construct compared to the world outside of school: (1) individual work and learning in school versus collaborative and cooperative learning outside;

(2) pure mental activities in school versus application and manipulation of activities outside; (3) symbol-based and abstract learning in school versus contextualized rationalization and reasoning outside; and (4) generalized, broad learning in school versus situation-specific skills and competences outside. Is it possible to recreate real world contexts in schools given the structures and constraints in schools? Unlike the real world, tasks created in schools are not authentic because students are not accountable for the decisions made (Cumming & Maxwell, 1999).

In a provocative article, Deborah Meier, founder of schools like Central Park in New York and Mission Hill in Boston, insists that "artificiality is …the raison d'etre for schools" but also acknowledges that "artificiality doesn't have to be a bad word, and authenticity isn't a guarantee of good education" (Meier, 1998, p. 596). Meier's (1998) resolution is "to create schools whose authenticity is self-evident." By this, she asserts that "authentic practices will follow [the] announcements of authentic purposes, not the other way around" (Meier, 1998, p. 605). This suggestion is shared by Splitter (2009, p. 143) who proffers that curriculum, teaching, and assessment strive to achieve "*what ought to be*" rather than attempt to establish a similarity between classroom activities and what goes on in the real world. To this end, teachers, schools, and educational systems have to be very certain of the types of learning they wish to measure and have to be able to ascertain why these are important before designing the assessments (Meier, 1998). An example is illustrated by the schools Meier founded. Each school is guided by an overarching philosophy centered on five critical *Habits of Mind*. Working with these broad goals, Meier and her teachers planned learning that ensures the relevance of the five habits in the daily lives of students, and the authentic assessments were aligned to these habits which were part of students' experiences in school.

Psychometric issues. Reliability, validity and accuracy are psychometric issues associated with authentic assessment (Gresham, 1991; Terwilliger, 1997). Gresham (1991) was concerned that authentic assessment advocates were more preoccupied with deliberations and discussions about the nature and characteristics of authentic assessments than with reliability, validity, and measurement error. In response, authentic assessment champions (e.g., Newmann, Brandt, et al., 1998) assert that psychometric standards were secondary because tests should first meet teaching standards and educator needs. For authentic assessment proponents, consequential validity or the implications for students and schools of using a particular type of assessment, is more appropriate. To this end, when authentic assessment is aligned to authentic instruction (Gulikers, et al., 2004), "any task can, in a way, be viewed as authentic if in our interpretation it's in keeping with its purpose" (Meier, 1998, p. 597).

Practical issues. Authentic assessments are time and cost intensive (Christenson, 1991) and require higher levels of teacher assessment literacy than are currently available (Cizek, 1997; Wolf, et al., 1991). In the stampede to embrace authentic assessment despite low levels of assessment literacy, overzealous teachers may end up using ineffective approaches to create an authentic task (Cumming & Maxwell, 1999). Designing authentic assessments is also demanding, as teachers must be able to devise learning goals, develop lessons that are integral to assessment, and interpret the data collected from the assessments (Brooks & Brooks, 1993; Fischer & King, 1996).

Finally, parents and teachers worry that if students spend too much time on the authentic assessments, they will not be adequately prepared for the standardized tests (Newmann, et al., 1995). However, research indicates that students exposed to authentic learning perform as well as, or better than peers prepared in the traditional drill-and-practice approach (Newmann, et al.,

1995). For example, research by Darling-Hammond, Ancess and Falk (1995) on the authentic assessments used by five restructuring schools reported that such assessments exerted more rigorous demands on students than traditional tests that focused on routine practices and facts.

Summary

Conceptually, authentic assessment appears to be still in its infancy because a large portion of the scholarly work is based on one source: Newmann and Associates' (1996) *authentic intellectual work*. The youthfulness of this concept is evident in that 'authenticity in teaching' continues to be a contested topic, and that the literature does not provide a definitive definition (Kreber, Klampfleitner, McCune, Bayne, & Knottenbelt, 2007). The philosophical, semantic, psychometric, and practical issues further attest to the nascence of this concept.

With respect to the semantic debate, while this dissertation acknowledges that the term, 'authentic' is contentious, the decision is to retain its use. This choice is made in relation to the characteristics of authentic intellectual work as conceptualized by Newmann and Associates' (1996) and Wiggins (1989), in particular, the fact that this type of assessment provides students with opportunities to be engaged in meaningful tasks or activities, and that require them to use the type of cognitive skills that they would in the world outside of school. This use is consistent with the way the Program for International Student Assessment (PISA), a triennial international study organized by the Organization for Economic Cooperation and Development (OECD) conceives of authentic contexts in the assessment of mathematics and scientific literacy. In the PISA 2009 Mathematics Assessment Framework (OECD, 2009), real world assessments or tasks involve

> moving beyond the kinds of situations and problems typically encountered in school classrooms. In real-world settings, citizens routinely face situations in which the quantitative or spatial reasoning or other cognitive mathematical competencies would help clarify, formulate or solve a problem. Such situations include shopping, travelling,

> cooking, dealing with personal finances judging political issues, etc. Such uses of mathematics are based on the skills learned and practiced through the kinds of problems that typically appear in school textbooks and classrooms. However, they also demand the ability to apply those skills in a less structured context, where the directions are not so clear, and where the student must make decisions about what knowledge may be relevant and how it might be usefully applied (OECD, 2009, p. 24)

The OECD's conception of real world tasks mirrors those envisaged in authentic intellectual work, in that there is scope for disciplined knowledge (skills learned and practiced in class and from textbooks), for construction of knowledge (basic skills applied in a less structured context) and for the value beyond school (skills used in situations that involve shopping, cooking, etc). In a sense, it is the demands made on the nature and quality of students' work that warrants authentic academic achievement (Hanley-Maxwell, et al., 1999). As a result, this dissertation retains the use of 'authentic' for the tasks, outcomes, and assessments. Although the term causes discomfort among some scholars, one way of resolving the issue is to "see [authentic] as merely an issue of label choice" (Ladwig, 1998, p. 118). While this does not settle the issue, it points to the need for further theory building, and perhaps, also more empirical work.

Empirical research on authentic intellectual work

This section presents the empirical work relating to authentic intellectual work (AIW) dating from the period 1990 to 2011. The year 1990 was selected as it is just after the publication of Archbald and Newmann's (1988) *Beyond standardized testing: Assessing authentic academic achievement in the secondary school* and Wiggins' (1989) *A true test: Towards more authentic and equitable assessment*. Both works are widely credited with the introduction of the 'authentic assessment' concept (Darling-Hammond, et al., 1995; Splitter, 2009; Tanner, 2001; Terwilliger, 1997). I limited the review to peer-reviewed journals because these have been selected for publication based on rigorous and methodologically sound studies. Drawing from the conceptual work on authentic assessment, I included published technical

reports, conference papers and other publications associated with the teams of university-based researchers from the Wisconsin Center of Education Research, the Consortium for Chicago School Research, the University of Queensland, the University of Newcastle (New South Wales), and the Center for Research Pedagogy and Practice. The majority of empirical studies applies or adapts Newmann's three criteria for authentic achievement. There were, however, other studies which analyzed authentic assessment without referencing Newmann. A few studies (n = 6) were selected because their conceptual description and definition of authentic assessment share the key features of Newmann's framework. This review only includes empirical work that studies teachers' classroom work with students. As such, empirical studies that involve electronic platforms that create a real world context for students are not included. The empirical work is presented according to the emergent categories based on the main research questions.

The review of the empirical work on authentic intellectual work is presented in four main sections, beginning with the introduction of seven studies (associated with the university-based research teams) conducted on teaching, instruction, and school reform to provide a broad overview of these research projects, their research objectives and methods, and the way in which they have applied Newmann's framework and methodology. The four sections documenting the research on authentic assessment are:

- Authentic intellectual work and the research on school reform (only descriptive and background to the large-scale university-based research studies that applied and adapted Newmann and Associates' (1996) framework)
- Authentic intellectual work and observation studies
- Authentic intellectual work and intervention studies
- Summary of conceptual and empirical work on authentic assessment

The empirical work, including the studies on reform, is categorized as **intervention** and **observation** studies. As the intervention studies emerge chronologically later, they are presented after the non-intervention studies. For both groups of studies, the common themes are the

examination of patterns and quality of teacher assessment and student work, exploration of the relationship between teacher assessment and student work, analysis of the impact of using authentic intellectual work on student achievement and other outcomes (e.g., engagement, perceptions, and motivation), and evaluation of the methods used for rating teacher assessment and student work. The findings of the studies that examined the methodology are important as they inform the research method this dissertation uses.

Authentic intellectual work and the research on school reform

In an in-depth study of classroom practices in the USA spanning a century, Cuban (1984) observed that despite a litany of school reform efforts, there has been little change in classroom practices – teachers typically do not alter their practices, preferring instead to adhere to traditional grammars of teaching and learning (Tyack & Tobin, 1994). Recently, the spotlight has turned on the classroom because "classroom practice is the heart of schooling" (Hayes, et al., 2006, p. 32). There are limitations to the research methods used to study classroom practices. Possibly the most direct means of documenting classroom practices would entail researchers sitting in, monitoring, and recording lesson proceedings. However, this method is both human and cost intensive to implement, and frequently only involves a small number of classrooms (Clare & Aschbacher, 2001). A cost effective way is to implement surveys but this method is constrained by the veracity of the self-reported teacher responses (Mayer, 1999). Consequently, one approach advanced involves the rating of teacher assessment and student work completed in response to these assessments (Clare, 2000; Clare & Aschbacher, 2001; Clare, Valdes, Pascal, & Steinberg, 2001; Matsumura & Pascal, 2003). This method was widely used in the study by the Center for Organization and Restructuring of Schools (CORS) which was set up in 1990 by the US Department of Education to embark on a five-year research study of school restructuring.

This section introduces seven research studies managed by university-based teams that employed or adapted Newmann and Associates' (1996) AIW criteria and standards. Like the work done by CORS, the examination of authentic intellectual work was conducted within the auspices of large-scale studies on school reform. Specifically, the researchers in these teams applied the criteria as indicators of the quality of instruction, assessment, and learning. The six studies which draw on Newmann's criteria and standards for authentic intellectual work are:

1. Fred Newmann and Associates from the Center on Organization and Restructuring of Schools (**CORS**) for the School Restructuring Study (**SRS**);
2. Fred Newmann and colleagues at the Consortium on Chicago School Research for the Chicago Annenberg Research Project (**CARP**)
3. Bruce King and colleagues at the Research Institute on Secondary Education Reform for Youth with Disabilities (**RISER**)
4. Bob Lingard and colleagues at the University of Queensland for the Queensland School Reform Longitudinal Study (**QSRLS**)
5. James Ladwig and colleagues from the University of Newcastle (Australia) for the Systemic Implications of Pedagogy and Achievement in New South Wales Public Schools Study (**SIPA**); and
6. Allan Luke and colleagues at the Center for Research in Pedagogy and Practice for the Core Research Program (**CRPP-CRP**).

The seventh research study on school reform is led by Lindsay Clare Matsumura from the National Center for Research on Evaluation, Standards, and Student Testing at the University of California, Los Angeles for the Los Angeles Annenberg Metropolitan Project (**LAAMP**). This study is not part of the list above. It is distinguished from the other six studies in that its theoretical framework is different, but its research focus and method draws on Newmann and Associates.

This section only introduces these research studies, their conceptualization and use of the AIW criteria, research objectives, and methods. The findings reported by these studies are integrated into the findings from the other studies included in this review that have similar research questions.

1. School Restructuring Study by the Center on Organization and Restructuring of Schools (CORS-SRS). This research examined the effects of school restructuring on student performance (Newmann, 1996). For the period from 1990 to 1995, researchers from CORS-SRS examined data from over 1500 elementary, middle, and high schools throughout the United States. The researchers conducted site studies at 44 schools in 16 states. The objective of CORS-SRS was to analyze how organizational features of schools could be changed to improve the intellectual and social achievement of students (Newmann, 1996). The researchers were guided by this overarching research question: Under what conditions can school restructuring promote authentic student achievement? Of the six supplementary research questions, the one pertaining to teacher assessment and student work was *How can schooling nurture authentic forms of student achievement?*

The CORS-SRS involved 24 restructured public schools in the USA, divided evenly among elementary, middle and high schools, located in 16 states and 22 districts. These schools served large student populations, averaging over 700 students (Newmann & Wehlage, 1995), and were mostly located in urban areas. The research was conducted from 1991 through 1994, with researchers studying each school for one year. The data sources included on-site observations and interviews with teachers, administrators, parents, policymakers, and students. The subjects included for study were mathematics and social studies, selected because these two subjects offered interesting contrasts (Newmann & Associates, 1996). The levels studied were Grades 4 and 5 for elementary school, Grades 7 and 8 for middle school, and Grades 9 and 10 for high school. In total, CORS-SRS examined practices in 130 classrooms, conducting up to four observation lessons per class (n = 556 lessons observed). The researchers also gathered two assessment tasks per teacher from the observation class, one collected in the fall, and one in the

spring. In total, 235 assessments were assembled. To evaluate the quality of student learning, the researchers collected a full set of student work per assessment submitted by the teacher. In total, over 2000 pieces of student work were examined. Additionally, students were required to complete a short questionnaire describing their experience with the task. Newmann and Associates (1996) adapted their 14 standards for evaluating authentic pedagogy and authentic student performance to in order to evaluate the quality of authentic intellectual work in these 24 schools.

2. Chicago Annenberg Research Project (CARP). Fred Newmann extended the methodology used in the CORS-SRS project to the CARP, which was conducted from 1996 to 1999. The study had three main research aims: to examine the nature of instruction in Chicago classrooms; to document whether the quality of instruction was improving with time; and to analyze key factors contributing to the improvements (Newmann, Bryk, & Nagaoka, 2001). The research involved 12 elementary schools during 1996-97, increased to 16 schools during 1998-1999 period. In total, the study involved 74 teachers and over 700 urban school students. The CARP study differed from the CORS-SRS project as classroom observations were not included in the research design. The similarity is that both studies applied the AIW criteria and standards to examine the quality of teacher assessment and student work.

3. Research Institute on Secondary Education Reform for Youth with Disabilities (RISER). Like CORS-SRS, the research and work emanating from RISER was grounded in school reform with a specific focus on students with disabilities. The overarching objective was to identify school reform practices that would benefit all students (King, et al., 2001). RISER's research drew extensively on Newmann's framework, constructivist theories and inclusive learning. The RISER philosophy rejected behaviorist theories of learning because

they entailed a deficit view of the learner, and as such, the learning opportunities for students with disabilities were consequently lowered (King, et al., 2001). Of the six core RISER research questions, the one that is germane to this study is: What are critical features of instruction, assessment, and support strategies that promote authentic understanding, and achievement (and performance for all students)? The research included in this review was conducted during the 1999-2000 school year, with the data coming from 32 teachers in 4 high schools in urban, rural, and small city contexts. Teacher assignment and student work were collected for Grades 9-12 English, mathematics, social studies, and science. In total, the work from 650 students was analyzed.

4. **Queensland School Reform Longitudinal Study (QSRLS)**. The QSRLS was the largest observational study of classroom pedagogy and student outcomes in Australian education in 1998 (Luke et al., 2000). The research sought to identify classroom practices that were most effective at producing positive academic and social student learning outcomes (Hayes, et al., 2006). In this way, QSRLS extended the conceptualization of student outcomes to include social outcomes. The study was conducted from 1998 to 2000 in Queensland government schools, and was led by two principal investigators, Bob Lingard (University of Queensland) and James Ladwig (University of Newcastle, New South Wales). The QSRLS research design included large-scale surveys of teachers and principals as well as case study analyses of 24 schools. Within these schools, the research team observed 975 lessons, and evaluated teacher assessments and student work from those lessons. Lessons were observed for English, mathematics, Studies of Society and Environment, science, and interdisciplinary lessons for Years 6, 9 and 11. Site visits were made twice a year to the case study schools. As did the

CORS-SRS researchers, the QSRLS team collected entire-class sets of student work samples from the participating teachers.

Conceptually, the QSRLS team made several significant modifications to the Newmann framework, beginning with the nomenclature. First, the Queensland scholars preferred to use 'pedagogies' over 'instruction,' because it captured teachers' professional practice, and had a "constructivist heritage" dating back to Vygotsky (Hayes, et al., 2006, p. 20). The second significant change was the use of the term 'productive' rather than 'authentic.' This nomenclature change was in response to the scholarly disagreement over the term, 'authentic' and its antonym, *inauthentic*. The Queensland team did not want to suggest that there was *real* or *true* student performance, pedagogy, or teacher assessment (Hayes, et al., 2006). The Australian scholars settled on 'productive' as a means to resist the market discourse that denounces and blames teachers for lackluster educational standards. Thus, in place of 'authentic instruction,' the Australian scholars used *Productive Pedagogies*.

The Queensland researchers made modifications to the Newmann standards. First, they divided the *pedagogies* criteria into four dimensions, namely, intellectual quality, connectedness, supportive classroom environment, and working with and valuing differences. The first two of these four dimensions are the same as those in Newmann's framework. The *Supportive Classroom Environment* dimension refers to the standards for instruction from an older iteration of the Newmann framework (See Table 2.3). The researchers included this dimension because student interviews indicated that the students valued positive relationships with their teachers (Hayes, et al., 2006).

Second, the QSRLS researchers redefined 'Authentic Performance.' The purpose was to expand the narrow focus on academic outcomes in Newmann and Associates' (1996) framework

and to align themselves to the Queensland curriculum and syllabus documents (QSRLS, 2001) in which learning was perceived as having both academic and social outcomes. Consequently, the Queensland researchers saw authentic achievement as having two subcategories: academic outcomes and social outcomes. They labeled these *Productive Performances* (Hayes, et al., 2006; QSRLS, 2001). Social outcomes require students to demonstrate their valuing and understanding of differences among cultural groups in Australia (cultural knowledges), display traits congruent with the responsibilities of being a citizen (responsible citizenship), and exercise responsibilities (transformative citizenship) (QSRLS, 2001).

The final set of modifications made by the QSRLS team involved the re-grouping of Newmann's 14 standards into new subcategories, and the inclusion of new ones. In total, the Queensland team developed 20 standards for productive pedagogies, 18 standards for productive assessments, and 8 standards for productive performance. One example of a change was the refinement of Newmann's *Construction of Knowledge* into *Problematic Knowledge* to indicate knowledge is not all objective, but is constructed in and dependent on the political, social, and cultural contexts (Hayes, et al., 2006).

In *Productive Assessment*, 'Problematic Knowledge' comprises construction of knowledge and consideration of alternatives. The Queensland team agreed with Newmann and Associates that assessments should demand higher-order thinking skills but extended this to require that students explore solutions, arguments, strategies or points of view when they examined a concept, problem, or issue (Hayes, et al., 2006). The documentation and explanation of the additions and modifications are found in Hayes, et al. (2006). Table 2.5 exhibits the key dimensions and standards used by the QSRLS scholars. The superscript numbers 1 through 4 illustrate the adaptations from the original Newmann framework and nomenclature. The

superscript alphabets ‘a’ and ‘b’ illustrate the standards that fall under academic and social performances respectively.

5. Systemic Implications of Pedagogy and Achievement Study (SIPA). James Ladwig, one of the two directors of the QSRLS, directed this study in Australia’s New South Wales public schools. SIPA’s overall design drew on the Queensland and Newmann models of authentic pedagogy (Ladwig, Smith, Gore, Amosa, & Griffiths, 2007), and had a teacher professional learning perspective. This longitudinal study examined the relationship between teachers’ professional learning, the quality of pedagogy, and student achievement for three overlapping cohorts of students over a period of four years. The schools (n=26; 15 primary and 11 secondary) were selected based on the stratified sampling. Once the school was selected, all students in Years 4, 6, and 8 were included in the study. Similar to the CORS-SRS and QSRLS projects, the SIPA study required teachers to submit their classroom assessments (n=78), and samples of student work (n=2236) in English and mathematics at different interviews of the school year. In total, the researchers examined 1374 pieces of student work.

Table 2.5

Adaptation and modification of Newmann & Associates' (1996) authentic achievement in Queensland

	Productive Pedagogies[1]	Productive Assessment[2]	Productive Performance[3] (Academic [a] and Social [b])
Intellectual Quality	Problematic knowledge	Problematic knowledge: Construction of knowledge[4]	Problematic knowledge [a]
		Problematic knowledge: Consideration of alternatives[4]	
	Higher-order thinking[4]	Higher-order thinking[4]	Higher-order thinking[4a]
	Depth of knowledge[4]	Depth of knowledge: Disciplinary content[4]	Depth of understanding[4a]
	Depth of student understanding[4]	Depth of knowledge: Disciplinary processes[4]	
	Substantive conversation[4]	Elaborated communication[4]	Elaborated communication[4a]
	Metalanguage[4]	Metalanguage	
Connectedness	Connectedness to world beyond the classroom[4]	Problem connected to the world beyond the classroom	Connectedness to the world beyond school[4b]
	Knowledge integration	Knowledge integration	
	Background knowledge	Link to background knowledge	
	Problem-based curriculum	Problem-based curriculum	
		Audience beyond school	
Supportive Classroom Environment	Social support[4]		
	Students' direction	Students' direction	
	Explicit quality performance criteria	Explicit quality performance criteria	
	Academic engagement		
	Student self-regulation		
Work with and Value Differences	Cultural knowledge	Cultural knowledge	Cultural knowledge
	Active citizenship	Active citizenship	Responsible citizenship
	Narrative	Narrative	Transformative citizenship
	Group identification in learning communities	Group identification in learning communities	
	Representation		

[1] Newmann's Authentic instruction
[2] Newmann's Authentic assessment
[3] Newmann's Authentic achievement, split into academic and social outcomes
[4] Newmann's standards (some were renamed)

[a] Academic performance
[b] Social performance

Adapted from Hayes, et al. (2006, pp. 22-23) and Gleeson (2011).

6. Center for Research in Pedagogy and Practice Core Research Program (CRPP-CRP).

Singapore's Center for Research in Pedagogy and Practice (CRPP) was set up in 2002 with the aim of developing comprehensive educational research programs that would provide evidence for evaluating Singapore's educational reforms and policies (Luke, et al., 2005; Luke & Hogan, 2006). The *TSLN* vision provided the impetus and context for the type of reform of which CRPP-CRP's research focused. Researchers developed the Core Research Program (CRP) which had six separate panels, each connected to and nested within each other in order to enable the triangulation of qualitative and quantitative data (Luke & Hogan, 2006). Of the six panel studies, the research questions for Panel 5 focused specifically on the nature of classroom assessment: What are the actual assessment practices that teachers put into practice in classrooms? How do these mediate and moderate intellectual and cognitive demand and depth? How are these linked to the quality of the written work that students produce in response to the assigned tasks? (Koh & Luke, 2009, p. 292). Similar to the Queensland study, the CRPP-CRP research made explicit use of discourse analysis and curriculum theory (Luke, et al., 2005). During the research period from 2004-2005, 59 schools (30 elementary and 29 high schools) participated in the study. The researchers collected Grade 5 and 9 English, social studies, mathematics, and science teachers' assessments (n=4097) (Koh & Luke, 2009).

Conceptually, the CRPP-CRP applied nine criteria to evaluate teachers' assessments and six for student work. The CRPP-CRP team adopted some criteria from the Newmann framework and devised others within the Classroom Coding Scheme. One significant difference was that the Singapore researchers did not attempt to measure 'higher-order thinking,' a feature that was used by the CORS-SRS and QSRLS projects. Instead, the Singapore researchers examined teachers' and students' representation of knowledge in the classroom (Koh & Luke, 2009). In

this way, the Singapore study shifted its focus from cognitive and thinking to textual representation (Koh & Luke, 2009). Table 2.6 shows the adaptations the Singapore team made to the Newmann framework.

Table 2.6
Adaptation and modification of Newmann & Associates' (1996) authentic achievement in Singapore

Newmann's Criteria	Teacher Assessment Standards	Student Work Standards
Construction of Knowledge	Knowledge criticism[N] Presentation of knowledge as given Comparison and contrast of knowledge Critique of knowledge Knowledge manipulation[Q] Reproduction Organization, interpretation, or evaluation of information Application / problem-solving Generation / construction of knowledge new to students Supportive task framing[C] Structure of the task Content scaffolding Procedural scaffolding Strategy scaffolding Clarity and organization Clarity and organization[N] Learner support Student control[C]	Knowledge criticism[N] Presentation of knowledge as given Comparison and contrast of knowledge Critique of knowledge Knowledge manipulation[Q] Reproduction Organization, interpretation, or evaluation of information Application / problem-solving Generation / construction of knowledge new to students
Disciplined Inquiry	Depth of knowledge[N] Factual knowledge Procedural knowledge Advanced concepts Sustained writing[N]	Depth of knowledge[N] Factual knowledge Procedural knowledge Advanced concepts Sustained writing[N]
Value beyond school	Connection to the real world beyond the classroom[N]	Connection to the real world beyond the classroom[N]

Adapted from Gleeson (2011), Koh and Luke (2009, p. 300 and 305 for standards to rate teacher assessment and student work respectively), and Koh, et al. (2005)
[N] Newmann's standard
[Q] QSRLS-added standard
[C] CRPP-CRP-added standard
Note: I have not attempted to fit the standard, 'Explicit performance standards / marking criteria' into the table. Although Koh and Luke (2009) applied this criterion when evaluating the quality of student work, it is more appropriate to view this criterion as a rating approach rather than a specific criterion for Authentic assessment or Authentic achievement.

7. Los Angeles Annenberg Metropolitan Project (LAAMP). This four-year study involved the evaluation of the Los Angeles Annenberg Metropolitan Program. The research team focused on developing and validating indicators to measure the quality of classroom practice. The areas differentiating the LAAMP research from the other six studies were its focus on the validity and reliability of the research methods. First, LAAMP researchers were interested in methodological issues like the reliability and independence of the teacher assessment scales, the relationship between classroom assignment ratings and other indicators of instructional quality, the number of assignments and raters needed to obtain a consistent estimate of the quality of classroom practice, the types of assessments, and the question as to whether the ratings of assignments and interviews provide similar estimates of practice. To this end, the findings reported by LAAMP provide useful guidelines as to the number of teacher assessments to collect, and the number of raters to employ to ensure that the estimates and measures are stable and accurate. In examining the rigor and feasibility of this research method, the LAAMP researchers also looked into the burden imposed on teachers who have to make arrangements for lesson observations, complete various research surveys, and make copies of student work.

Second, the LAAMP researchers did not draw explicitly on Newmann's AIW criteria and standards, but for developed their own rating rubric for classroom instruction (cognitively challenging and meaningful instruction, clear goals for student learning, substantive and specific feedback) and teacher assessment (cognitive challenge, clarity of the learning goals focused on student learning, clarity of the grading criteria, alignment of goals and tasks, alignment of goals and grading criteria, and overall quality). For both measures, the focus was on the level of cognitive challenge posed by teacher assessments.

Third, the LAAMP study (1998-2001) was smaller in scale than the other studies, and only focused on Language Arts lessons. In the first two years, 8 schools (4 elementary and 4 middle schools) serving mostly poor and minority students, and 24 teachers (evenly distributed between elementary and middle school) participated in the study. The levels of interest were Grades 3 and 7. Teachers submitted four language arts assignments, based on what they professionally judged to be *typical* and *challenging* writing, comprehension, and content-area writing assignments. This requirement for typical and challenging assignments was similar to the CARP study. In addition, for each assignment, teachers submitted four samples of student work, judged to represent medium and high quality. Teachers also completed a one-page summary for each of the four assignments and returned this to the research team. To determine the quality of classroom instruction, researchers observed lessons twice a year. The selection of the observation lessons was done purposefully, and principals were invited to suggest the dates and times for these visits. The researchers interviewed teachers during each site visit.

To examine the external validity of the findings from the first two years, the LAAMP researchers modified the scale and nature of the study to compare student achievement in schools serving predominantly Latino and African American students with those serving primarily middle-class students (mostly White and Asian). Another change was made to the research design. Instead of collecting four teacher assignments, the research team reduced this to two. This was to determine if collecting two (rather than four) teacher assignments would yield a stable and reliable estimate - if it did, this procedure would reduce the burden on teachers. Second, the researchers changed the type of schools in order to examine if there were differences in the quality of teacher instruction and assessment, and of student work among schools serving specific socioeconomic student populations.

Summary. As can be seen from the studies on reform, the research work that utilizes the criteria and vision of authentic achievement does not involve interventions or experimental designs. The common features among these studies are that they are all embedded in larger studies that seek to examine the nature of school reform through a number of indicators, of which one is the quality of teaching and learning. The research studies apply Newmann's AIW concept as indicators for rating the quality of teacher assessment and student work. While there have been adaptations and modifications to the nomenclature (Tables 2.5 and 2.6), the basic features of Newmann's Authentic Intellectual Work (Construction of Knowledge, Disciplined Inquiry, and Value Beyond School) are retained in the later work undertaken by the Australian and Singapore teams. However, the variations among the different models makes comparison of the quality of teacher assessment and student work across studies difficult as there is no common yardstick (Gleeson, 2011).

A second common feature of all the studies is that they adopt similar research designs to collect data on classroom instruction, teacher assessment, and student work. There were some variations in the type and nature of teacher assessments required: the CORS-SRS project asked teachers to submit assessments that they perceived would best elicit students' understanding of the subject (Newmann & Associates, 1996; Newmann, et al., 1996), the QSRLS collected teacher assessments that were considered 'best practice' (QSRLS, 2001), the CARP project requested teachers to submit samples of 'challenging' (defined as indicating that students understood the subject at a high level) and 'typical' (reflecting daily work) tasks (Bryk, et al., 2000). The data sources included lesson observations and their documentation over a period of time at each school site, as well as the collection of teacher assignments and student work.

The research analyses also adopted similar approaches to rate teacher assessments and student work. This involved teams of researchers and specially-recruited teachers who attended training sessions to acquire the language and skills needed in order to rate the collected artifacts. Ensuring rigor during the rating of the artifacts was a methodological practice that was reported in all the studies. This preview of the large-scale studies does not highlight the findings or the analysis methods used because they are presented in the sections below. Finally, a summary of the name and dates, samples, subjects, grade levels, and research methods are shown in Table 2.7.

Table 2.7

Summary of the six studies[a] *on reform that adopt or adapt Newmann and Associates' (1996) framework*

Study name and date	Sample: Schools, classes, teachers, and students	Subjects Grade level	Lessons observed	Artifacts collected	Measure of authentic achievement
Center on Organization and Restructuring of Schools School Restructuring Study (CORS-SRS), 1990-1994[b]	24 schools (8 elementary, 8 middle, 8 high schools) 130 classes 2128 students	Mathematics and social studies Grades 4-5, 7-8, 9-10	n=132	234 teacher assessments 5100 student work	
Chicago Annenberg Research Project (CARP), 1996-1997[b]	12 elementary schools (pilot) and 16 (final) 74 teachers 700 students	Language arts (reading and writing) Mathematics Grades 3, 6, and 8	No	349 teacher assessments 3300 student work	Authentic Intellectual Work
Research Institute on Secondary Education Reform for Youth with Disabilities (RISER), 1999-2003[b]	4 high schools 32 teachers 32 classes 650 students	Language arts, mathematics, science, and social studies Grades 9-12	Yes	51 teacher assessments 314 student work	
Queensland School Reform Longitudinal Study (QSRLS), 1998-2000	24 schools (12 primary and 12 secondary) 250 teachers	English, mathematics, science, Studies of Society and Environment Years 6, 9 and 11	n=975	321 teacher assessments 2450 student work	Productive Pedagogies, Assessment & Performance
Systemic Implications of Pedagogy and Achievement Study (SIPA), 2004-2007	26 schools (15 primary and 11 secondary) 1374 students	English and mathematics Years 4, 6, and 8	No	78 teacher assessments 2236 student work	Authentic Intellectual Work
Center for Research in Pedagogy and Practice Core Research Program (CRPP-CRP), 2004-2005	59 schools (30 elementary, 29 high schools)	English, social studies, mathematics and science Grades 5 and 9	Yes	4097 teacher assessments and associated student work	Adapted Authentic Intellectual Work and Singapore Classroom Coding Scheme

[a] The LAAMP study is not included in this table because the studies reviewed were concerned with methodology rather than evaluating large-scale school reform.

[b] Adapted from Newmann, King & Carmichael (2007, p. 17)

Authentic intellectual work and observation studies

The international studies (n=25) presented in this section of the review are observation studies because they do not involve the AIW concept as an intervention or an experiment. Rather, these studies examine existing practices and school reform by employing the AIW criteria and standards fully or adapting them so as to measure the quality of teaching practices and learning. A common thread that runs through these studies is the focus on patterns in the quality of teacher assessment (n=11) and student work (n=13), the correlation between the quality of teacher assessment and student work (n=8), and the measuring of the impact of authentic intellectual work on student achievement (n=8). A cluster of observational studies (n=3) were also dedicated to determining the reliability and accuracy of the number of teacher assignments to collect, the number of raters, and the cohesiveness of the rating scale. Finally, there are studies that examine teachers' professional practice and growth (n=3), the value of authentic assessment at a system level (n=1), and the nature of authenticity (n=2). A recent dissertation (Gleeson, 2011) was included in this review because of its unique research design enabling the tracking of the quality of assessment practices of novice teachers from their teacher preparation course to their third year of teaching. This allows for an examination of whether teachers' practices change over time. The classification of the studies was based on the main research questions.

Quality of authentic intellectual work in teacher assessment. The researchers interested in examining the quality of teacher assessment adopted similar research methods to answer this question. Specifically, the research method involved a specially trained team of researchers and teachers using a rubric to rate teacher assessments. Newmann and Associates' (1996) pioneering research method involved teachers in the participating schools submitting two assessments they considered valid and important indicators of their students' understanding of

the subject (Newmann & Associates, 1996; Newmann, et al., 1996). A rating scale (a three- or four-point Likert scale) was assigned to each of the AIW standards. Teachers also responded to a survey to provide details of the task. For each submitted task, teachers included a set of student work. The scoring of the artifacts was conducted by researchers and teachers currently teaching the subject. To ensure rigorous standards in the rating process, the two raters discussed differences in rating until a consensus was met. The internal consistency for the rating was .79 (Cronbach's alpha) (Newmann, et al., 1996). The QSRLS, CRPP-CRP, RISER, CARP and SIPA studies employed the same process as CORS-SRS, as did a small study examining authentic achievement in Title Schools (see D'Agostino, 1996).

Broadly, the findings from the six large-scale studies indicated that there were huge variations in the quality of teacher assignments, even among schools identified for their innovative practices after the reform process. Across the 24 schools they studied, Newmann and Associates (1996) reported that 60% of the variability in quality of teacher assessment was found within individual schools while 40% of the variability was found between schools. This is a strong indicator of the tremendous variations in the quality of assessments among teachers. The variations in schools could be explained by differences in students' socioeconomic backgrounds as reported by the QSRLS team in Australia. They observed that assessments in schools serving students from higher socioeconomic backgrounds had higher expectations of students.

While the scholars reported instances in which they found exemplars of high quality teacher assessments, the majority of assessments received low ratings. For example, in the CORS-SRS analyses, out of a total score of 43 (high intellectual demand) and a minimum of 11 for authentic pedagogy (instruction and assessment), the highest score was 33.5 and the lowest was 12.5 (Newmann & Associates, 1996). Reporting similar patterns in Queensland, the QSRLS

team expressed that they were "disturbed by the low intellectual demand of the assessment task" (Lingard, et al., 2006, p. 90).

The varied nature of teacher assignments was also reported in the aforementioned small-scale dissertation study (Gleeson, 2011). This study examined as to whether the quality of teacher assessments created by a cohort of beginning teachers (n=11), who attended a teacher education program committed to the principles of social justice and student learning, changed with more teaching experience. Gleeson's research method mirrored that of the studies on reform. Teacher tasks (n=53) and student work were collected and rated using the RISER (King, et al., 2001) standards. The mean ratings for each teacher over three years indicated that the quality of the tasks varied across the group of teachers (Gleeson, 2011). For some, there was no change in the quality of assessments created and used over the three year period. For other teachers, the ratings had a large range in the mean scores. Overall, Gleeson concluded that there was no overall pattern in teachers' assessment tasks. One reason for the absence of any statistically significant differences in the ratings of teacher assessments is the small sample size, which did not meet the statistical assumptions of the repeated-measures ANOVA analyses which Gleeson used to test the differences between the mean scores.

The variation of teacher assessment was reported to vary by academic subject (Gleeson, 2011; Koh & Luke, 2009; Newmann, Lopez, et al., 1998; QSRLS, 2001). This could be attributed to the nature of the subject. In Singapore, teacher assessments for social studies scored higher ratings than for mathematics, science, and languages (Koh, et al., 2005), while in Queensland and Singapore, science assessment tasks were found to be more diverse and intellectually challenging (Koh & Luke, 2009; QSRLS, 2001). In the CRPP-CRP study, social studies assessments at both Grades 5 and 9 were found to be more intellectually challenging and

had statistically higher scores for criteria like 'connections to the real world beyond the classroom,' 'sustained writing,' and 'student control' (Koh & Luke, 2009). One reason for the higher authentic quality of social studies assessments was that it is a non-examination subject at Grade 5 (Koh & Luke, 2009). In the rating of the assessments used or created by eleven novice teachers in Gleeson's (2011) study, social studies and writing tasks also received the highest mean scores. This suggests that the design of these assessments provided the most authentic intellectual challenge for students. Like the researchers in the other studies, Gleeson (2011) suggests that this is due to the nature of the subject. For instance, writing tasks probably rated high on the AIW criteria, because the assignments frequently required students to write extensively. Furthermore, the nature of the writing tasks contributed to higher ratings for the criterion, 'connection to students' lives.'

The quality of teachers' assessment varies by grade level (Gleeson, 2011; Ladwig, et al., 2007; Newmann, Lopez, et al., 1998; QSRLS, 2001). Newmann, Lopez, and Bryk (1998) reported that Grade 6 and 8 teachers in Chicago created more mathematics and writing assessments that were in the 'moderate' and 'extensive' authentic challenge categories than Grade 3 teachers. Ladwig et al. (2007) reported that in the SIPA study in New South Wales, the scores for secondary tasks were higher than their primary counterparts. Newmann, Lopez and Bryk (1998) found that Grade 6 writing assignments were more challenging than those for Grades 3 and 8. In Australia, the QSRLS team also found that easier tasks were assigned to Year 8 than Year 6 students (Lingard, et al., 2006). These findings indicated that teachers either underestimated their students or did not challenge them sufficiently.

The nature of teacher assessments varied according to student background, in particular, students' track or program (King, et al., 2001; Koh, et al., 2005), and the type of school they

attended (Clare, et al., 2001; Gleeson, 2011). In the RISER study, 35 teachers submitted work from two students in their classes, one with a disability, and one without a disability. Teachers also had to submit a checklist of accommodations they had made, where necessary, for both groups of students. Based on the ratings, the RISER researchers reported that because of accommodations, the AIW rating for tasks designed for students without disabilities had an overall higher mean score than the tasks intended for students with disabilities. Though small, the difference in mean scores between the two tasks was statistically significant. The LAAMP researchers (Clare, et al., 2001) also reported differences in AIW that teachers assigned between students attending schools serving mostly poor Latino and African American communities and those attending schools with a majority of White and Asian students. The students from the latter schools also had higher scores when the Stanford Achievement Test was used as an indicator of prior achievement. Based on the analyses of student work, Clare et al. (2001) found that teachers' assessments from the higher achieving school rated significantly higher in several standards, including level of cognitive challenge, clarity of the learning goals, alignment of the goals with the tasks, and overall quality. This suggests that students attending higher achieving schools were given more intellectually challenging tasks than their counterparts attending schools that serve low income minority populations. The findings from the LAAMP study raise questions as to the assignment of classes to teachers. Were the better teachers sent to teach higher achieving students, or were teachers committed to providing more opportunities to students whom they perceived had the caliber to handle more difficult tasks?

Finally, the source of teacher assessment was found to be a variable associated with the quality of authentic intellectual work. The mean scores of teachers who created their own assessment tasks were higher on the AIW scale than the scores for teachers who did not create

their own assessments (Gleeson, 2011). This difference in the mean scores was statistically significant, despite the small sample of teachers in Glesson's (2001) study (n=22). It suggests that novice teachers in the study were capable of providing challenging tasks for their students, and points to the value of teachers creating their own tasks rather than relying on commercially produced assessments. Teacher-created tasks are likely to be more responsive to curricular, teaching and learning goals as compared to those that are constructed for the mass market.

Overall, the scores from the rating of teacher assessments were not encouraging. The findings reported in the empirical work indicated that based on the scores assigned to the tasks, teachers had low expectations of students. Broadly, the analyses of the mean scores on each criterion indicated that teachers still operated within the behaviorist paradigm because the researchers repeatedly found that teacher assessment required students to recall and reproduce knowledge. In Singapore, the CRPP-CRP researchers reported a preponderance of high scores in the categories for 'factual knowledge,' 'procedural knowledge,' and 'presentation of knowledge as given/truth, and reproduction' (Koh, et al., 2005). The Queensland team found a number of tasks that only required students to demonstrate their note-taking abilities (Lingard, et al., 2006), leading the researchers to conclude that Queensland teachers were merely occupying students with 'busy work' (QSRLS, 2001, p. 27). Designing tasks that encapsulated the *Value Beyond School* criterion was found to be most lacking in the teacher assessments (King, et al., 2001; QSRLS, 2001). Among the tasks collected by the RISER team, most tasks scored reasonably in the *Construction of Knowledge* criterion, and for the 'elaborated written communication' standard. However, these tasks scored dismally for the *Value Beyond School* criterion. One reason for this finding is that there are difficulties in designing tasks requiring students to address real world problems, or to make connections across disciplinary areas (King, et al., 2001).

In summary, through the collection and analysis of teacher assessment, the empirical research reported that there were large variations in the authentic intellectual work quality of teacher assessment. Based on the rating scales the researchers used, the mean scores of teacher assessment fell within the lower range. On the whole, across the different studies, there are variations by

(a) School – The quality of teacher assessment was higher in schools serving students from higher social economic or high achieving backgrounds. This suggests that in such schools, there are higher expectations of students.

(b) Subject – The variations in the quality of teacher assessment are associated with the nature of the subject. The empirical work reported higher teacher assessment scores for subjects like social studies and writing which were found to be intellectually more challenging.

(c) Grade level – Typically, the higher the grade, the higher the quality of teacher assessment. Yet in Chicago and Queensland, the researchers found more challenging tasks being designed for lower grades rather than for the higher ones. This suggests that teachers are either underestimating or not sufficiently challenging their students.

(d) Student background – The quality of teacher assessment varied according to the track, program, or school in which students were placed. The mean scores for teacher assessments were lower for students with disabilities, students studying in schools serving low income populations, and students placed in less demanding tracks.

(e) Origin of assessment – Teacher-created assessments were assigned higher mean AIW scores than those obtained from commercial producers.

Finally, when teacher assessment was examined in relation to specific AIW standards, the empirical work reported that teacher assessments required students to recall and reproduce facts and knowledge instead of challenging them with higher-order tasks such as making connections or applying classroom knowledge to real world problems.

Quality of authentic intellectual student work. The research examining the quality of authentic intellectual student work involved the rating of collected artifacts, using or adapting Newmann and Associates' (2006) AIW standards. However, while teacher assessment was rated for *Value Beyond School*, the rating of student work did not involve this criterion. The standards for the *Value Beyond School* criterion for the CORS-SRS were not developed because of logistic constraints (Newmann & Associates, 1996). The researchers lacked resources to interview students in order to ascertain if the tasks they completed were meaningful or valuable beyond school. Thus, the only measure for this criterion was in teacher assessment.

The procedure for rating student work is the same as described above for teacher assessment. However, there were some variations in the criteria pertaining to the collection of student work. For example, the CORS-SRS and QSRLS teams collected student work for an entire class, the SIPA team collected work for an entire year group over six time periods, and the CRPP-CRP researchers collected four samples each of high, medium, and low quality student work. Although the LAAMP researchers asked for student work that represented different ability levels, they only focused on student work from the high and medium ability levels. This was because their focus was to understand how teachers intended their assessments to be completed by students, and not to obtain a distribution of student work in a particular classroom.

The research presented the quality of student work descriptively and statistically. In a number of studies, exemplars of student work were displayed with annotations to illustrate how a

particular response reflected the demands stipulated in each of the rating criteria (see for example Clare & Aschbacher, 2001; Koh & Luke, 2009; Newmann, et al., 1996). Typically, examples illustrating high, medium and low quality student work were presented. The quality of student work was analyzed statistically, using means and standard deviations as indicators of the distribution of and variations in quality.

As with the authentic intellectual quality of teacher assessment, there was variability in student work (Newmann & Associates, 1996). Overall, the quality of student work indicated moderate (Gleeson, 2011) or lackluster levels of performance (Koh, et al., 2005; QSRLS, 2001), based on the criteria used by the research teams and on the analyses of the data. Specifically, the evidence indicates that quality of student work fell below the theoretical mean (QSRLS, 2001). Gleeson's (2011) analyses indicate that the mean scores for student work were lower than the midpoint on the scale. The QSRLS researchers attributed the quality of student work to the nature of the tasks assigned by teachers. They noted that there were exemplars of high quality student work, but these were not demanded by the task. In fact, the QSRLS researchers suggest that "many [teacher] tasks tended to limit the students to low levels of performance" (QSRLS, 2001, p. 22), a similar conclusion reported by Newmann et al. (1998).

There was a subject effect in the pattern of students' work (Gleeson, 2011; Koh, et al., 2005; Koh & Luke, 2009; Newmann, et al., 1996) as rated using the AIW criteria. Typically, student work in social studies and writing scored higher ratings than work in other subjects (Gleeson, 2011; Newmann & Associates, 1996). Research analyzing student work in social studies and writing indicated that students were able to analyze information and elaborate on their understanding of content (Gleeson, 2011). Comparatively, the rating for student work in

mathematics and science showed that students were primarily involved in recalling information, and there were few opportunities for them to apply knowledge to new situations (Gleeson, 2011).

When the rating scores were further disaggregated, there were variations in the criteria for each subject. For example, the CRPP-CRP team found that student work in Grade 5 social studies scored higher than other disciplines for standards like 'advanced concepts,' 'critique of knowledge,' and 'evaluation of information.' Comparatively, student work for Grade 5 science scored high means in areas such as 'compare and contrast knowledge,' and 'application/problem-solving.' In terms of the distributional pattern, Newmann, et al. (1998) found that at least 17 percent of student writing was categorized as 'extensive intellectual work,' meaning that the quality of work involved students constructing knowledge; demonstrating mastery of grammar, vocabulary, and usage for the specific grade level (i.e., disciplined inquiry); and producing elaborated written communication. Comparatively for mathematics, just less than 2 percent of student work was classified as 'extensive intellectual work,' based on the rating of the completed tasks. The findings from these three research teams indicate that differences in the quality of student work may be associated with the nature of the discipline. The CRPP-CPP team found that there were more opportunities for students to engage in performance-based tasks (e.g., role-play) that were closely aligned to the world outside school in social studies than in other subjects (Koh & Luke, 2009).

The authentic intellectual quality of student work varied by grade level (Koh, et al., 2005; Ladwig, et al., 2007). In general, raters assigned a higher rating to student work at the higher grade. For instance, the CRPP-CRPP reported that Singapore's Grade 9 students had a higher rating for 'Advanced Concepts' as compared to younger students. Surprisingly, while student work for social studies scored the highest for standards like 'critique of knowledge,' and

'interpretation,' the mean ratings for Grade 9 student work for social studies were lower than the Grade 5 student work for 'compare and contrast,' and 'connections beyond the classroom.' In New South Wales, the SIPA study reported a similar pattern. Notably, the team found that in Year 8, the quality of student work for mathematics was lower than that in Year 4. The same finding was reported for Year 8 Human Society and Its Environment (HSIE), which had a lower mean than Year 4. However, Ladwig et al. (2007) are cautious about this finding because the scores pertain to different cohorts of students. The results reported by SIPA and CRPP-CRP need to be interpreted cautiously because the researchers do not point out if the differences are statistically significant.

The pattern of authentic intellectual student work varied according to student background with variations by academic track (Koh, et al., 2005), school attended (Gleeson, 2011) and student status (King, et al., 2001). This evidence suggests that the quality of authentic intellectual work produced by students from less privileged backgrounds lags behind that of their privileged peers. Here, 'privilege' refers to students attending more advantaged schools (e.g., urban vs. suburban), placed on more prestigious academic tracks, and not identified as special needs. Researchers from the CRPP-CRP team reported variations in the quality of student work according to the educational track. In Singapore students in the more demanding streams (i.e., EM1, EM2, Express) scored consistently higher in all the CRPP-CRP's standards than did students in the less demanding courses [i.e., EM3, Normal (Academic), and Normal (Technical)]. This pattern was the same for both Grade 5 and Grade 9. However, the researchers did not report if the differences in these scores were statistically significant. At the same time, students from the EM3 track had a higher rating for 'Connections to the real world' than their counterparts in the EM1 and EM2 tracks. Students studying in suburban schools produced higher quality work

than their peers in urban schools. In Gleeson's (2011) study, despite the disproportionately larger number of students attending urban schools, students from suburban schools scored higher AIW means than their peers in urban schools. This difference in the means was statistically significant and is noteworthy because of the small sample size in Gleeson's study.

There are differences in the quality of authentic work between students with disabilities and their non-disabled peers. King, et al. (2001) reported that the mean rating of work for students without disabilities was higher than the mean rating for students with disabilities. This finding was statistically significant. However, the RISER researchers also found that the work of 62% of students with disabilities was at the same, or higher, level of authenticity than their matched non-disabled peers. Furthermore, of 13 students with disabilities who scored lower than their peers, 4 of them were given tasks that scored lower in authenticity too.

The authentic intellectual work quality of student work varied according to the origin of the tasks that they were assigned (Gleeson, 2011). The mean scores on the authentic achievement criteria in Gleeson's (2011) were statistically higher for students whose teachers created their own assessments as compared to students whose teachers used variations of standardized tests or commercially-produced assessments. As with the higher means reported for teacher-created assessments, this finding indicates the value of teacher-designed assessment.

Two other studies examined the quality of authentic intellectual student work, using research methods that differ from the research teams in the six large-scale studies on school reform. The findings from the first study illustrate that students have difficulty coping with more authentically challenging work. Bishop (2000) examined gifted students' thought processes and feelings as they engaged with an authentic research project. This is a small-scale, qualitative study conducted among ten junior high gifted students tasked to complete independent research

projects. Using multiple data sources (observations, interviews with students, the teacher, and librarians, and documentary sources) the researcher took on the role of a participant observer, spending 12 weeks in the classroom and documenting the entire research process that students underwent. Overall, the quality of student thought and work in the project was wanting (Bishop, 2000). This was evident in the presentations in which students provided simple explanations instead of demonstrating 'elaborated communication' as envisaged by Newmann and Associates (1996). Although this study involved a small sample of students in gifted education, its findings are valuable in identifying the difficulties students face when assigned to complete authentic tasks independently. In addition, the study's findings contributed to the debate on 'authenticity' because despite giving students the leeway to initiate projects of interest, the study found that they did not find the tasks or learning meaningful or engaging.

Regardless of age or stage in education, students tended to perform better in tasks requiring them to reproduce and recreate knowledge rather than to produce and create new information. This was evident in a study of the quality of undergraduates' writing. This Scottish study analyzed 100 undergraduates' essays focused solely on the *Disciplined Inquiry* criterion and its related standards, namely analysis, elaborated written communication, and disciplinary concepts. Macllelan (2004) chose to focus purely on this criterion because it resonated with Facione's (1990) notion of 'critical thinking skills.' To this end, the aim of the study was to describe the tenets of disciplined inquiry as found in the undergraduates' essay responses. This study was not an intervention as the essay prompts provided for the undergraduates were part of an educational psychology module on motivation. Each undergraduate had to respond to three (out of five) essays. In terms of the analysis of student work, the essays were scored in a way similar to the CORS study, in that there were several raters (university professors) and

established procedures for inter-rater reliability (Maclellan, 2004). Each essay was rated on a 3-point scale based on the standards for 'disciplined inquiry.' Statistical procedures (chi-squared tests, Friedman tests and the Wilcoxon Test) were employed to examine the statistical significance of differences among the essay scores, and across the standards. Interestingly, the findings on the quality of undergraduate student work were similar to those of the elementary and middle school students reported earlier. The undergraduate essays were found to be lacking in the 'analysis' standard with most of them seeing knowledge as "non-contestable" (Maclellan, 2004, p. 79).

The rating of student work by researchers only provides one perspective of the quality of authentic intellectual learning in the classroom. An important question is whether teachers' perception of student performance matched the ratings assigned by the researchers and teacher-rater teams (Aschbacher, 1999). In this aspect, the findings reported in the research were mixed. Correlation analyses indicated low to moderate levels of correlations between teachers and raters, and the ratings showed that teachers tended to 'overrate' student work (Aschbacher, 1999). One study sought to understand teachers' perception and interpretation of student work by examining whether the ways teachers discussed their goals and their understanding of assessment and student learning reflected the AIW concept (Gleeson, 2011). The findings indicated that teachers whose objectives and conceptions of assessment and learning aligned with authentic intellectual work had a higher tendency to use assessment tasks that received high AIW scores (Gleeson, 2011). Conversely, teachers who did not speak about authentic assessment, but focused on lower-order skills, were more likely to develop or use assessments with very low levels of authentic intellectual work. These findings suggest that teachers first have to be cognizant of the importance of authentic assessment in order to use assessments that engage students in these

competencies (Gleeson, 2011). To this end, "teachers' beliefs about teaching and learning influenced their teaching practices and the decisions they make related to assessment" (Gleeson, 2011, p. 303). Research from this perspective is valuable in the examination of the quality of classroom practices. It examines whether teachers' understanding and perception of high quality work resonates with that of researchers. If teachers are consistently overestimating the quality of work that they provide, then t implications for practice need to be addressed.

In summary, while there are variations in the quality of students' authentic intellectual work, the findings presented in this section suggests that throughout all levels of education – elementary through undergraduate – and in a variety of international settings (e.g., the USA, Australia, Singapore, and Scotland), the quality of student work is not high. On the whole, student work was reported to be deficient in the areas valued as critical and necessary for life in the 21st century. Overall, the research on the quality of student work found variations by

(a) Subject – The ratings for the quality of student work were higher for subjects like social studies and writing. This is attributed to the nature of the subjects which provided opportunities for students to engage in discussion and extended communication.

(b) Grade level – The higher the grade, the higher the rating of student work. However, the research findings indicated that even at lower grades, student work could receive high ratings for the AIW standards.

(c) Student background – There were variations in the quality of student work by academic track, school attended, and student status. Overall, the findings indicated that the quality of student work was lower for students from disadvantaged backgrounds.

(d) Origin of task – The work of students whose teachers created their own assessments received higher ratings compared to students whose teachers used commercially produced assessments.

One significant finding indicated that when compared to the researchers' ratings of student work, interviews with teachers suggested they tended to overestimate the quality of their students' work. At the same time, the rating of teacher assessment (under 'Quality of authentic intellectual work in teacher assessment') indicated that teachers' underestimated students' abilities, and hence, the research reported that teacher assessments did not commensurate with the grade level. In view of this misalignment, and given the emphasis many societies place on higher-order thinking skills, these findings point to the need for professional development in order to prepare teachers to design challenging tasks for their students.

Relationship between teacher assessment and student work. This review has presented the ratings of the quality of teacher assessment and of student work. The question is whether there is an association between the quality of teacher assessment and the quality of student responses. To examine this, correlation analyses were conducted between the ratings for teacher assessment and those for student work (Clare & Aschbacher, 2001; King, et al., 2001; Koh & Luke, 2009; Newmann, Lopez, et al., 1998; QSRLS, 2001). These analyses indicated a significant positive correlation between teacher assessment and student work (Clare & Aschbacher, 2001; King, et al., 2001; QSRLS, 2001). The findings suggest that task demands rated lower in authenticity were associated with student work which was also rated lower in authenticity, and vice versa.

The quality of tasks that teachers assign is correlated to the quality of the responses that students construct and provide. More bluntly, "what you test is what you get" (Koh, Lee, Gong,

& Wong, 2006, p. 6). The CRPP-CRP team found that the majority of pairs of correlations were statistically significant and had moderate to large correlations. Putting this in the context of the pattern of teacher assessment found in the CRPP-CRP study, Koh and Luke (2009) expressed concerns about the quality of student learning in Singapore. Given that the greater proportion of teacher assessments scored high on lower-order criteria like knowledge reproduction and factual recall, this suggests that there is a low threshold or expectation set for students. The question as to whether students were sufficiently challenged and provided with opportunities to display higher-order thinking and complete more complex tasks. In addition, the moderate to strong correlation between teacher assessment and student work suggests that what teachers assign to students signals what matters, "what will count" (Koh & Luke, 2009, p. 312).

In exploring the relationship between teacher assessment and student work for CARP, Newmann, et al. (1998) compared the features of student work in response to 'typical' and 'challenging' assessments. They reported that while students were able to complete the 'typical' assessments, the rating of the work according to Newmann's criteria and the examination of the responses indicated that they did not produce authentic intellectual work. In comparison, high-scoring student work in both mathematics and writing had features of authentic intellectual work. The researchers concluded that the deficiency on the part of the 'typical' student work was due to the demands of teacher assessments, and not student ability. They surmised that

> if an assessment makes low demands for authentic intellectual work, students will almost surely score low on the standards for authentic performance, because they will have virtually no opportunity to show proficiency in construction of knowledge and disciplined inquiry (Newmann, Lopez, et al., 1998, p. 35).

Newmann et al. (1998) tested this proposition by employing a two-level Hierarchical Linear Model to examine the structural relationship between the quality of teacher assessments and the average student work produced in those classrooms. The magnitude of the correlations

was "impressive," – the values ranged from 0.7 for Grade 6 to a high of 1.0 for Grade 3 mathematics. This result indicates that "the quality of the assessments teachers assign … is virtually deterministic of the quality of work that students produce on average" (Newmann, Lopez, et al., 1998, p. 50).

The findings on the statistical correlation between teacher assessments and student work are significant in that they support authentic work proponents' assertions that students, regardless of background, are capable of handling complex and intellectually challenging tasks (Archbald & Newmann, 1988; Newmann & Associates, 1996; Newmann, Brandt, et al., 1998; Wiggins, 1989, 1990). Authentic achievement and assessment have been criticized for neglecting basic skills and knowledge (e.g., Terwilliger, 1997, 1998). Wiggins' (1989) vision for students to demonstrate mastery equivalent to experts in the field has been criticized as being narrow in terms of educational goals. Opponents of authentic achievement value disciplinary knowledge, such as that expounded in E.D. Hirsch's (1987) *Cultural literacy*, as important educational goals (Terwilliger, 1997). They take issue with the rigor, value, and purpose of authentic assessment because they are concerned that authentic intellectual work, especially with its emphasis on *Value Beyond School*, is achieved by sacrificing important knowledge for skills.

However, the findings from the empirical work contradict these reservations. They show that when students are exposed to authentic intellectual work, they are able to perform as well as, or even better than their peers who participate in a behaviorist drill-and-practice learning environment (Newmann, et al., 1995). In particular, King et al.'s (2001) research findings were compelling because they showed the high quality work that students with disabilities are capable of, given the appropriate accommodations. This finding also illustrates clearly how Vygotsky's *zone of proximal development* comes into play within constructivist theories. As a result, despite

teachers assigning tasks pitched at a higher level, their provision of the appropriate level and amount of supports and accommodations will enable students to demonstrate their mastery of complex learning tasks. The findings reported in this review send a powerful signal to teachers to adopt high expectations for all students – an essential tenet of constructivist learning theories.

Effect of authentic intellectual work on student achievement. One of the underlying assumptions of authentic intellectual work is that it is more likely to "motivate and sustain students in the hard work that learning requires" (Newmann & Associates, 1996, p. 9). In addition to reporting patterns in the quality of teacher assessments and student work, the research examining authentic intellectual work also sought to explore its effects on student learning.

The studies selected for this review used statistical analyses to examine the relationship between authentic intellectual work and academic achievement as represented by student performance on standardized tests (e.g., Ladwig, et al., 2007; Newmann, Bryk, et al., 2001), as well asby student engagement (D'Agostino, 1996; Marks, 1995). Therefore student learning transcends learning of academic content to include experiences in school that are part of 'affective engagement' (Fredricks, Blumenfeld, & Paris, 2004). Within the studies on student academic achievement, the research provided a status report of the effect of authentic academic achievement (e.g., Newmann & Associates, 1996), or examined a longitudinal change over time (V. E. Lee, et al., 1995, 1997).

The methods for examining the effect of authentic intellectual work on students involve the use of regression models, and in particular, multi-level or Hierarchical Linear Models (HLM) because of the nested nature of the data – students nested in classrooms, and classrooms nested in schools (Bryk, et al., 2000; Newmann, Bryk, et al., 2001). The basic statistical model used authentic intellectual work as an independent variable with scores from standardized tests as the

dependent variable. Further analyses proceeded by controlling for other variables of interest to policy makers. These variables are pertinent at the student (minority status, SES, gender, and prior achievement) and school (minority concentration, type of school, and school SES) levels.

To examine the impact of authentic pedagogy on student achievement, the CORS-SRS researchers regressed teachers' authentic pedagogy scores against student scores on the AIW standards using a three-level Hierarchical Linear Model (Newmann & Associates, 1996). Overall, the effects were positive, and the effect of employing authentic pedagogy was 0.37. This indicates that students taught by a teacher using authentic pedagogy were predicted to perform 0.37 units higher if the teacher taught and assessed using the three AIW criteria.

From this basic regression model, further analyses were conducted by controlling for other variables, such as students' social (specifically, gender, race, ethnicity, socioeconomic status) and academic (performance at the National Assessment of Educational Progress) backgrounds. In terms of the level of schooling, the CORS-SRS researchers reported that the effect of authentic pedagogy was highest for high school followed by middle and elementary school respectively (Newmann, et al., 1996). However, there was some variation in the pattern because the effect of authentic pedagogy in mathematics was high in elementary and high schools but lower in middle schools. The project team suggested that this variation might be attributed to the small sample of schools, which did not allow for sufficient statistical power in the analyses.

With regards to equity issues, researchers examined whether authentic pedagogy promoted high achievement for all students regardless of their background and status (e.g., Newmann & Associates, 1996; Newmann, Bryk, et al., 2001). Based on the HLM computation using the research procedure described above, the CORS-SRS and CARP teams concluded that

irrespective of gender, race, ethnicity, and SES, authentic pedagogy exerted the same effects on performance in all classes. In the analyses, CARP researchers used student achievement from two norm-referenced tests, the Iowa Tests of Basic Skills (ITBS) to measure achievement in reading and mathematics, and the Illinois Goal Assessment Program (IGAP) to measure student achievement in reading, mathematics, and writing. The statistical method used was HLM. Based on the analyses, Newmann, Bryk, and Nagaoka (2001) declared that after controlling for race, SES, gender, and prior achievement differences among classrooms, there were substantial benefits when students were given tasks aligned to authentic intellectual work in mathematics and writing – the gains were 20 percent greater than the national average. More importantly, the team also found that regardless of high or low prior achievement, all students appeared to benefit almost equally from being assigned authentic assessments. As such, the team concluded that "in short, authentic intellectual assignments enrich instruction not only for able children, but for all students" (Newmann, Bryk, et al., 2001, p. 23)

The CORS-SRS team also found that while Hispanic and low-SES students did not achieve significantly lower scores than White or high-SES students, African American students did not perform as well as white students, and female students outscored male students significantly. The researchers compared these gaps to the achievement gap on a traditional standardized test, and concluded that the inequality in authentic performance is no greater, and possibly less than the inequality as measured on a traditional test (Newmann, et al., 1996).

The impact of using authentic pedagogy varied by subject. Within the Chicago schools context, a study conducted during the 1993-1994 academic year reported that the impact of authentic pedagogy varied by subject (D'Agostino, 1996). In this study of 53 classrooms in 29 schools, D'Agostino (1996) found that while authentic instruction in mathematics was

significantly and positively associated with gains in achievement, the reverse was true for reading since instruction was not a significant predictor of mean classroom gains in vocabulary. One possible reason for this finding was that D'Agostino only observed one lesson each for mathematics and reading, as compared to multiple lessons by the CORS-SRS group.

Students did not always have equal access to or benefit equally from authentic intellectual work. Scholars in the SIPA study in New South Wales, Australia, reported a different pattern from the American scholars. However, it must be noted that there were slight differences between the CORS-SRS and SIPA analyses. While CORS-SRS used a three-level hierarchical model, the SIPA model only had two levels. Both studies used different programs to estimate the multi-level models: CORS-SRS used HLM while SIPA ran their analyses on the MLWIN program. Statistical analyses for the SIPA model indicated that the effect of authentic pedagogy was in favor of female students and of those with higher prior achievement. In terms of the equity issue, the effects were not in favor of students from aboriginal, low SES, and English speaking backgrounds. Furthermore, the quality of the authentic task had a lower effect on achievement for SIPA than for CORS-SRS. In estimating the effect of authentic work, the SIPA researchers rationalized that their findings differed from the CORS-SRS findings possibly because their measure of 'authentic instruction' only included teacher tasks, whereas the CORS-SRS measure included both lesson observation and teacher tasks. Additionally, the use of a two-level statistical model could have muted the analyses because it underestimates level two effects.

Researchers also examined whether the gains from authentic pedagogy were sustained over time (V. E. Lee & Smith, 1994, 1995; V. E. Lee, et al., 1995, 1997). Lee and colleagues conducted observation studies on the longitudinal impact of authentic pedagogy. The study was based on surveys and testing data from the 1988-1992 National Educational Longitudinal Study

(NELS) database which tracked over 10,000 students in 1,000 American schools from 8th through 12th Grade. The surveys for mathematics and science instruction included items that were consistent with Newmann and Associates' (1996) AIW criteria. These were used as estimates of the degree of authentic intellectual demands that students experienced. Lee and colleagues (1995, 1997) used HLM to model the relationship between authentic work and student achievement from 8th to 10th Grade and 10th to 12th Grade, controlling for student SES and prior achievement. At the school level, the researchers controlled for average school SES, minority concentration, school sector, and school size (V. E. Lee, et al., 1997). The quality of authentic practices was measured by the frequency of use as provided by the responses in the self-report surveys. The practices were organized into three groups, namely traditional, moderate, and restructuring practices (V. E. Lee, et al., 1995), and the growth trajectory discussed as 'early gain' (8th – 10th Grade) and 'late gain' (10th – 12th Grade). The findings indicated that the achievement gains of students attending restructured schools were significantly higher than the gains for those in traditional schools, and these gains were also more equitably distributed (V. E. Lee & Smith, 1994). In terms of the growth trajectory, the gains from attending schools employing higher levels of authentic pedagogy were sustained over the four years. However, the gains for both mathematics and science were higher in the early period than in the later period. The higher and sustained achievement of students attending schools that adopt practices aligned to authentic pedagogy provide support for the value of this approach to teaching, learning, and performance.

Students who experience authentic pedagogy become more engaged in learning because authentic intellectual work brings meaning to learning beyond providing the right answer or demonstrating competence (Newmann & Associates, 1996). This assertion was confirmed by

studies on the association between student engagement and authentic work (Marks, 1995). Researching the restructuring schools movement in the USA in the 1990s, Marks (1995) drew on survey data from the main CORS-SRS project that sought to elicit student responses about their attitudes, behaviors, and experiences in mathematics or social studies classes. In total, Marks used data from 3,660 Grade 5, 8 and 10 students in 143 of the 149 CORS-SRS classrooms, and analyzed student responses using two-way analysis of variance (ANOVA) and HLM. Based on the analyses, Marks (1995, p. 28) reported that "authentic academic work exerted a powerful influence on engagement." This is possibly due to the higher intellectual demands made on students (Marks, 1995).

There were variations in the level of engagement due to the use of authentic intellectual work: female students were more engaged than male students; elementary students were the most engaged and high school students were the least; and social studies and mathematics lessons were engaging for middle school students (Marks, 1995). A significant finding was that there was no difference in the level of engagement when the data was disaggregated by social class and race-ethnicity. This finding is congruent with research reporting the impact that restructuring schools are having on equity and student learning (Marks, 1995).

In summary, using authentic pedagogy was found to have a positive impact on the scores of student work and student engagement. After controlling for race, social economic status, gender and prior achievement, there were benefits when students were given tasks aligned to the AIW standards. However, the distribution of the benefits differed among student sub-groups. In one study, Hispanic and low-SES students in the USA achieved significantly lower than white or high-SES students (Newmann & Associates, 1996) while in another study, the gains in student

achievement were more equally distributed (V. E. Lee & Smith, 1994). Finally, the effects of using authentic pedagogy were found to be sustained over time.

Research method and analyses in studies on authentic intellectual work. The research method used by Newmann and Associates (1996), and subsequently the other research teams (e.g., CARP, QSRLS, LAAMP, and CRPP-CRP), was a novel way of examining classroom practices. This method involved three generic criteria (*Construction of Knowledge*, *Disciplined Inquiry*, and *Value Beyond School*), each with its corresponding set of standards to rate instruction, assessment, and student work. This section is a review of the empirical work on the processes to ensure reliability, accuracy, and consistency in the rating of teacher assessments and student work. It also provides a summary of the various analyses procedures used by the different research teams.

To ensure reliability, consistency, and accuracy in the rating of teacher assessment and student work, the research teams undertook comprehensive preparation and training of teacher raters (Clare & Aschbacher, 2001; Koh, et al., 2005; Koh & Luke, 2009; Newmann & Associates, 1996; Newmann, et al., 1996), provided anchor papers for training (Koh, et al., 2005; Koh & Luke, 2009), and instituted iterative rating sessions, involving double scoring of all or a randomly selected portion of the teacher tasks and student work (King, et al., 2001; Koh, et al., 2005; Koh & Luke, 2009; Ladwig, et al., 2007; Newmann, 1996; Newmann & Archbald, 1992).

There were similarities and differences in the scoring processes used to rate instruction, teacher assessment and student work. Generally, the teacher-raters taught the subject, while the trainers from the research teams had expertise in the subject (Newmann & Associates, 1996; Newmann, et al., 1996). For some studies, the rater would rate all three criteria for the same teacher task or student work (e.g., Newmann & Associates, 1996; Newmann, et al., 1996). In

other studies, raters were randomly assigned to different standards such that each piece of assignment was scored by different people (Bryk, et al., 2000). The purpose of the latter method was to control for rater bias. To ensure consistency and reliability among the raters, the research teams conducted periodic checks on the amount of exact and adjacent-point agreement. These statistics were documented and reported, as an indication of rigor and reliability in the research process.

Where there was a difference in scores between the raters during the rating process, there would be negotiations until the discrepancy was resolved (King, et al., 2001; Koh, et al., 2005; Koh & Luke, 2009). If the score differences were more than two points, the training team member made the final decision to award the point (Newmann, Lopez, et al., 1998). Typically, the studies employed a minimum of two raters, sometimes including a third rater who would double-code a set of randomly assigned student work. Teacher work was always double-rated, but not student work, due to the sheer numbers. Reliability of the scoring process was reported in three ways: percent of exact agreement, percent of agreement for adjacent scores (difference of one score point), and correlation coefficient. With the rigorous procedures, and dedicated training, the percent exact agreement between two scorers ranged from 70 to 80%, and the percent exact agreement for adjunct-scores ranged from 88 to 100%.

Other than reporting the reliability and accuracy of the process, researchers (e.g., Clare, 2000; Clare & Aschbacher, 2001; Clare, et al., 2001; Matsumura & Pascal, 2003) also examined the number of teacher assignments and raters needed to obtain a consistent estimate of the quality of classroom practice, determine the reliability and independence of the teacher assignment rating scale, explore the relationship between teacher assessment ratings and other indicators (i.e., student work), and examine the extent to which the ratings of teacher assessment

were aligned to teachers' interview responses on their practices. The significance of these five studies by the LAAMP researchers lies in their review of the methodology and the ensuing analyses which guide the design of this dissertation.

To establish the number of teacher assignments and raters needed to obtain a consistent and stable estimate of the quality of classroom practice, the LAAMP researchers utilized a decision study to estimate generalizability coefficient of different numbers of raters and assignments (Clare, 2000; Clare & Aschbacher, 2001; Clare, et al., 2001). They conducted a generalizability study to investigate the quality of the design in terms of obtaining consistent estimates of classroom practice. Using three raters and four assignments, they obtained a G-coefficient which indicated that the variation in the ratings was due to differences across teachers rather than raters. The statistical test indicated that a minimum number of two raters were needed for each school level, i.e., two raters for the four elementary schools, and two raters for the four middle schools in the study.

Subsequently, the LAAMP researchers explored the use of two teacher assignments instead of four. This was an attempt to reduce the burden on teachers who spoke about the time constraints faced when required to submit four assignments to the team. The computation from the decision study indicated that two teacher assessments were insufficient to obtain a "stable estimate of quality" (Clare, et al., 2001, p. 30). As such, they recommended that for a small sample (n=30) of teachers, it would be necessary to employ two raters and to collect four assessments. However, the researchers were also mindful that teachers had heavy professional loads, which should be considered when determining the number of assessments to collect from them. In a subsequent study with schools in Los Angeles, Matsumura and Pascal (2003) examined whether the use of three assignments would yield a stable estimate of quality. The

computation from the generalizability study yielded a stable estimate of quality at the secondary but not at the elementary level. The researchers surmised that this difference might be attributed to the fact that elementary teachers submitted a mixture of commercially-produced as well as teacher-created assignments. Comparatively, secondary teachers only submitted teacher-created assignments (Matsumura & Pascal, 2003). Based on these findings, the LAAMP researchers recommended collecting a minimum of two assignments per teacher.

To estimate the reliability of the teacher assessment rating scale, Cohen's kappa coefficient was used to estimate the proportion of agreement between raters after chance agreement was removed, and also the inter-rater reliability (Clare, 2000; Clare & Aschbacher, 2001; Clare, et al., 2001). The latter refers to the degree to which the raters were able to independently examine the same assessment and decide on a score. They reported that the kappa coefficients for each of the 'typical' and 'challenging' assessment types were statistically significant. The coefficients of Cronbach's Alpha, used to estimate the internal consistency of the ratings, were within a statistically acceptable range, varying from .68 to .91. The LAAMP researchers also reported a high percentage of exact agreement among the raters (greater than 80%). These estimates were used as guidelines for this dissertation.

The LAAMP researchers examined the relationship between the teacher assessment ratings with student work (Clare, 2000; Clare & Aschbacher, 2001). They conducted a correlational analysis of the dimensions of the teacher assessment scale and of those in the student writing scale. They reported that there was a statistically significant association between teacher assessment and student work. However, they also point out that their statistical test only sought to identify if there were an association between the two dimensions. They did not conduct

a test for direction of influence between the quality of teacher assessments and the quality of student work.

Examining conceptions of 'authentic' assessment. Under the conceptual analyses of 'authentic assessment,' I presented the scholarly discourse that questioned the meaning of this type of assessment. Some philosophical scholars (e.g.,Splitter, 2009) argue that authenticity is relative – what is authentic for one person may not be authentic for another. Deborah Meier (1998) questioned whether schools in themselves were authentic, given the curriculum structures and block schedules. Within the empirical work, two studies examined 'authenticity' in a teaching context (Rahm, Miller, Hartley, & Moore, 2003), and in a standardized assessment (Grant, Gradwell, & Cimbricz, 2004). In both studies, researchers deliberated the meaning of the term, and how it might be understood within the context of the study.

Taking up the debate over 'authentic' context in education, Rahm, Miller, Hartley, and Moore (2003) drew on the collaboration among students, teachers, and scientists in two case studies to support their assertion that authenticity is not a static concept, but is relative and contextualized within the interactions and interdependences of the scientists, teachers, and students in the case studies. The researchers used interviews and field notes to document the discussions between scientists and teachers regarding the development and use of a plot of land. Through analyses of the data, the researchers illustrated how authentic science education emerged, and how it became authentic for the purpose of the group using the land. For instance, teachers worked in a variety of contexts, with different students, and had specific concerns and research interests for their students. Authenticity in this case study is about the interactions and interdependences. To this end, Rahm et al. (2003) assert that authenticity needs to be seen as an "emergent" concept that is "diverse in meaning" (p.737).

Standardized tests are widely used because they are easy to score, have validated psychometric advantages, and are easier to design. However, they are not authentic because they assess basic knowledge and skills in a decontextualized way that does not resemble how such knowledge is used in the world beyond school, and convey to students that there is always a correct response. As an answer, Grant, Gradwell, and Cimbricz (2004) examined whether large-scale assessments can be authentic. This empirical study examined whether the document-based question (DBQ) in the New York State examination is authentic to history and the activities that historians engage in. Unlike the ubiquitous multiple choice item, the DBQ included an essay prompt and constructed-response questions based on a set of seven or eight documents. The researchers examined the three tasks that appeared in the Global Exam in 2000, 2001 and 2002, and compared them to the curriculum framework. The analyses of the DBQ drew on characteristics of authentic assessment created by Jay McTighe, Grant Wiggins, and Fred Newmann and Associates. The analyses indicated a number of areas of dissonance between the DBQ and the nature of historians' work: historians collaborate rather than work individually when examining sources, and the editing of the sources rendered the DBQ inauthentic – it now reflects 'authenticity' in relation to the editor's perspective. As a result, while a commendable effort, the DBQ was only superficially authentic in capturing the work of historicans (Grant, et al., 2004). The researchers questioned whether it was possible to develop an authentic task within a large-scale assessment, especially when an authentic task would require collaborative work and oral communication. The researchers rationalized that authentic teaching and assessing might only be pertinent to classroom work, and not feasible for large-scale testing.

These two studies contribute significantly to advancing the work on authentic assessment. They push for deeper examination of the meaning of 'authenticity' in teaching and assessment.

Second, the findings from both studies speak to the difficulties and challenges of designing and developing authentic instruction, contexts, and assessments. Together, these studies suggest that authentic assessment is an emerging field and warrants further work.

Authentic intellectual work and intervention studies

Included in this part of the review are 12 empirical pieces on authentic intellectual work that involve an **intervention**, for instance, specifically designing an assignment that is aligned to Newmann and Associates' (1996) AIW framework (e.g., Avery, 1999; Gulikers, et al., 2004), creating a professional program on authentic intellectual work (e.g., Dennis & O'Hair, 2010), and developing authentic contexts for assessment and learning (e.g., Doering & Veletsianos, 2008; van't Hooft, 2005). With the exception of three studies, the empirical research on authentic intellectual work is recent (e.g., Dennis & O'Hair, 2010; Manning, Sisserson, Jolliffe, Buenrostro, & Jackson, 2008). Eight of these studies apply Newmann and Associates' (1996) AIW criteria. Of these studies, all except one (Ben-Chiam, Keret, & Ilany, 2007), were conducted in the USA. In comparison to the observational studies reviewed above, especially those emanating from the research centers, the studies in this section are mainly on a small-scale. The currency of the intervention work suggests a still emerging field. As with the observation studies, the review of the empirical research on authentic intellectual work that involves interventions on teachers' practice is categorized by the main research question.

Quality of authentic intellectual work in teacher assessment. Interest in teacher classroom practices has spurred interventionist approaches to improve the quality of teacher assessment. Researchers have devised professional development methods to enhance teachers' capacity in using authentic instruction and assessment (Avery, Freeman, & Carmichael-Tanaka, 2002; Dennis & O'Hair, 2010; Manning, et al., 2008). The professional development methods

include peer coaching (Avery, et al., 2002), developing a joint rubric (Manning, et al., 2008), and mentoring (Dennis & O'Hair, 2010). The professional development programs in the three studies presented in this section applied Newmann and Associates' standards. In one study, five teachers from three American high schools (one alternative school, one charter school, and one traditional school) who were active and interested in the Authentic Teaching Alliance worked with researchers to create and implement authentic tasks for their classrooms (Dennis & O'Hair, 2010). Using Newmann and Associates' framework to rate the teacher assessments, Dennis and O'Hair (2010) reported that the highest quality assessments were developed by the teachers in the charter school and the lowest quality authentic assessments originated from the teachers in the traditional school. From interviews with teachers, as well as analysis of teachers' portfolios and journals, Dennis and O'Hair (2010) concluded that the ideal setting that enables teachers to enact high quality authentic intellectual work is a small school with small classes, adequate administrative support, extra planning time for teachers, and appropriate funding.

A peer coaching program is one way to improve the quality of teachers' authentic instruction. One study tracked changes in the quality of teachers' (n=15) authentic instruction before and after the professional development sessions (Avery, et al., 2002). The researchers collected teacher and student work and rated them to establish the quality of authentic intellectual work after teachers completed the professional development sessions. Using pre- and post-test scores, Avery et al. (2002) analyzed the ratings using paired and independent t-tests to calculate the effect sizes. They reported that after professional development, there were improvements in teachers' instruction scores in secondary social studies in two areas, 'deep knowledge' and 'higher-order thinking.' The researchers reported statistically significant, medium effect sizes for authentic instruction. While this finding speaks to the value of

professional development programs as a means to improve the quality of teacher assessment, the small sample limits its generalizability to other contexts.

Acquainting teachers with authentic instruction frameworks is one way to increase the quality of teacher assessment. In Chicago, 128 teachers from 23 schools attended professional development sessions to develop rubrics based on Newmann and Associates' (1998) framework to assess the authenticity and intellectual quality of teacher assessment and student work in mathematics, English, science, social studies, and technology (Manning, et al., 2008). These sessions involved researchers modeling and guiding teachers on how to score and analyze the assignments. After analyzing the teachers' work, the researchers concluded that the mean scores for most teachers' assignments were still below what the researchers perceived as *good*. The weakness of this study is that the researchers did not present the scores in any form of ranking, or include any statistical analyses.

In summary, the evidence from these three studies point to the benefits of using professional development as an approach to improve the quality of teacher assessments. However, these studies do not report whether the benefits of the practices are sustained over time. It is also difficult to ascertain the effects of the intervention because the studies did not involve a comparison group.

Quality of authentic intellectual student work. Researchers working on interventionist studies explored the quality of student work following the intervention. Specifically, researchers examined the quality of student work at a moment in time (Manning, et al., 2008) or as change over time following the intervention (Avery, et al., 2002; Purcell-Gates, Duke, & Martineau, 2007). The studies also differ in that some examined the quality of student work based on their teachers undergoing professional development (Avery, et al., 2002; Manning, et al., 2008), other

studies specifically designed 'authentic' platforms to examine the quality of student learning and work (van't Hooft, 2005). Scholars also studied the quality of student learning based on the explicit teaching of genre features in science (Purcell-Gates, et al., 2007). Finally, the empirical work also sought to investigate students' perceptions about their learning (Doering & Veletsianos, 2008; van't Hooft, 2005) after being introduced to specific 'authentic' learning experiences.

Researchers reported mixed results in terms of the quality of student work following teachers' participation in specific professional development sessions. After working with teachers to develop and apply rubrics based on Newmann and Associates' framework, Manning et al. (2008) found that the mean scores of student work across the five subject areas in the study were lower than what the researchers would consider as 'good.' Comparatively, after teachers from a large inner-city school district in the USA attended a professional development session involving peer coaching to enhance their abilities to design and implement authentic intellectual work, Avery et al. (2002) reported encouraging outcomes – the analyses indicated that there were tremendous gains in the quality of student work. The effect sizes ranged from medium to large. While this finding is encouraging, the design of the study raised questions regarding student growth. As this study lacked a comparison group, it was difficult to attribute the portion of these gains that were due to students' development over the six-month period and the portion linked to the intervention (Avery, et al., 2002).

Students benefit when their teachers make deliberate and concerted efforts to apply the AIW standards. In fact, the performance and learning of students whose teachers explicitly planned, taught and focused on authentic reading and writing in science improved at a faster rate than their peers taught in classes with less focus on authenticity (Purcell-Gates, et al., 2007). In

this experimental study, the 16 teachers of 420 randomly selected students attended summer workshops on the intervention strategy and agreed to teach science twice a week. While the conception of authenticity in this study did not draw directly on Newmann and Associates' framework, the researchers applied a similar definition of authenticity, in that a literacy event that has an authentic purpose or function serves a "social communicative purpose" (Purcell-Gates, et al., 2007, p. 14). This feature is similar to the "substantive conversations" standard in Newmann and Associates' framework. Researchers observed lessons and coded the implementation of the intervention. Student learning was measured according to assessments the researchers designed. Student growth, performance, and fidelity to conditions were analyzed quantitatively using a variety of statistical analyses (including Hierarchical Linear Models), and student scores were produced using Item Response Theory approaches. The findings indicated that the authenticity of instruction mattered more than background variables. In fact, children from low SES homes developed at the same rate as their more affluent peers (Purcell-Gates, et al., 2007).

The quality of students' authentic learning may be improved through the creation and development of tools or authentic contexts which include technology-infused projects (van't Hooft, 2005) and the use of real-time authentic geospatial data to examine students' perceptions of learning science and geography (Doering & Veletsianos, 2008). Doering and Veletsianos (2008) drew on constructivist learning theories in their work with 65 Caucasian middle-school students in completing geospatial lessons. These lessons were introduced as an alternative to text-book sources of geographical data and learning, and the goal was to explore students' learning in geography. Using the constant comparative method to analyze interviews with students, Doering and Veletsianos reported that the use of real-time authentic data benefited

student learning in geography: the tool deepened students' understanding and motivated them to explore more geographic locations. Students also reported that they learned to collaborate with their peers. In the same manner, van't Hooft (2005) examined 99 Grade 7 students' perception of learning science after participating in technology-infused projects such as the Ohio Schools Going Solar (OSGS) project. The two-week OSGS project is designed to create an authentic learning experience. The project in the participating school required students to conduct research on alternative energy during two separate time periods. This task design resonated with Newmann's and Wiggins' conceptions of authentic assessment, as it required students to apply higher-order skills in a situation that mirrors real life. The task required students to work in teams to build a device powered by an alternative energy source. Teachers taught the basic content using traditional whole group instruction, and assigned individual student research. Van't Hooft employed pre- and post-surveys, the design of which were based on Newmann and Associates' (1996) authentic achievement criteria to elicit students' perceptions following their participation in the OSGS project. The data collected was analyzed using reliability coefficients. The analyses indicated that there was a positive effect in regards to students' perceptions towards the 'disciplined inquiry' and 'construction of meaning' constructs after having participated in the OSGS project (van't Hooft, 2005).

From these two studies, it is evident that despite the complexity of the tasks, students again rose to the challenge. As with the observation studies, this finding confirms the assertion that, when teachers assign high quality authentic assessment, the corresponding student work and engagement levels are also high.

Relationship between teacher assessment and student work. Similar to the observational studies, scholars who designed the intervention studies examined the relationship

between teacher assignments and student work produced in response to these assignments. Using correlational analyses, researchers (Avery, 1999; Manning, et al., 2008) working on the intervention studies found that the higher the teacher assessment score, the higher the corresponding student work completed in response to the task. Avery's (1999) study was significant because it addressed one of the weakness in the research design of Newmann and Associates' (1996) method, specifically, the variability in the interpretation of student performance. This is due to variety of teacher assessment tasks submitted in Newmann and Associates' (1996) study. To this end, students were not taught the same curriculum (Newmann, et al., 1996). This made it difficult to compare the quality of student work across schools and classrooms. Avery's (1999) study addressed this shortcoming by involving five high school history teachers working together to design an authentic assessment and instruction unit on immigration. The researchers observed and rated teachers' instruction according to Newmann's criteria. They also had students complete an engagement survey. Correlational analyzes reported a strong statistically significant relationship between teacher instruction and student work (Avery, 1999).While Avery's research addressed one weakness in the CORS-SRS, one drawback of this study is that it did not provide details as to how the five teachers were recruited for the study. In addition the small sample limits the generalizability.

Effects of authentic intellectual work on student achievement. There was one study which examined the effect of authentic intellectual work on student achievement following an intervention or experimental research study. Using student performance on a specially designed culminating project as the outcome variable, Avery (1999) regressed student scores on student demographics, student engagement, and authentic instruction. Together the three sets of independent variables explained 54 percent of the variance in student performance. However

when disaggregated, authentic instruction accounted for the largest proportion (40 percent) of the variability in scores, as compared to student engagement (7 percent) and student background (1 percent). The finding provided support for authentic assessment advocates' claims that the complex skills required to produce high quality authentic intellectual work would improve performance on traditional tests. Avery (1999) found that there was a statistically significant positive correlation between performance on the culminating project task and students' scores on a ten-item traditional multiple choice test. Both findings speak to the value of teachers spending time to develop complex higher-order thinking skills during instruction, while also providing opportunities for students to respond to challenging assessment tasks.

Examining conception of 'authentic' assessment. Researchers added to the extant epistemology and discourse in attempts to conceptualize 'authentic' assessment. One study examined the notion of 'authentic assessment' by developing a five-dimensional framework (task, physical context, social context, assessment result or form, and criteria and standards) for authentic assessment, and designed assessment tasks that comprised all or some of the dimensions of this framework (Gulikers, et al., 2004). The objective of the study was to explore if the five dimensions comprehensively encapsulated 'authenticity' in order to examine the relative importance of the five dimensions, and to compare teachers' and students' notions of the authenticity dimensions. Gulikers et al. (2004) assigned 28 nursing students to these tasks, and tracked their perceptions of the tasks assigned to them. Some students were assigned tasks that had all the five-dimensions while other students completed tasks that only had some of the dimensions. Teachers also discussed their views about the assessments. Students ranked *task*, *result or form*, and *criterion* as the most critical dimensions for an authentic assessment, and identified the dimensions, s*ocial context* and *physical context* as least important. Comparatively,

teachers perceived the *physical context* of the task to be the most important. Based on the findings, Gulikers et al. (2004) concluded that authenticity is a "multifaceted concept" (p.83) and stressed the value of including student perceptions in designing effective authentic assessments. The finding points yet again to a complex field that requires further scholarship to develop and understand the notion of 'authenticity.'

Authentic intellectual work and teacher learning. Specially-designed professional development helps teachers improve their design and use of authentic assessment. The studies (n=5) presented in this section focus on teachers' learning after attending professional development sessions designed around authentic assessment. Of the four studies, one applied Newmann's framework (Avery, et al., 2001), one adapted Newmann's framework (Koh, 2011b; Koh, Tan, & Ng, 2012), while two other studies (Ben-Chiam, et al., 2007; Grisham-Brown, Hallam, & Pretti-Frontczak, 2008) adopted a conceptualization of 'authentic' assessment or tasks that had a similar meaning. Israeli scholars Ben-Chiam, Keret, and Ilany (2007) defined an authentic assessment as one that is "genuine, trustworthy and usually presents problems with relation to everyday life" (p. 335). Their operational definition is significant as it indicates a converging common language being adopted for 'authentic' tasks, since scholars from a wider geographical area are also applying the same conceptualization.

A common thread across the findings is that dedicated professional development involved using specific teaching protocols (e.g., three dominant features in 'Promote LINKages,' Grisham-Brown, et al., 2008) to increase teachers' awareness of and confidence in using authentic assessment. Ben-Chiam et al. (2007) created proportional reasoning authentic investigative tasks and evaluated their impact on mathematical content, pedagogical knowledge, and attitudes of 15 pre-service elementary and middle school mathematics teachers in Israel.

The professional development leader modeled the strategies for presenting these tasks to young students. Based on pre- and post-session surveys documenting their attitudes, and their development in proportional reasoning, the findings indicated that the authentic task sessions resulted in an improvement among pre-service teachers in their attitudes towards teaching mathematics. Additionally, teachers were more confident in dealing with the topic of proportional reasoning (Ben-Chiam, et al., 2007).

Professional development sessions improve teacher competencies by introducing a common language when discussing authentic instruction. Avery et al. (2001) examined 16 Minnesota teachers' learning from attending a monthly Authentic Pedagogy in the social studies (APSS) seminar. The APSS sessions provided teachers with a common language and collegial support to discuss authentic instruction and develop the practice thereof. During the seminars, researchers and teachers developed assessments based on Newmann and Associates' framework, provided peer-critique of the tasks, and discussed ways to improve the tasks. At the implementation stage, teachers videotaped lessons and discussed the quality of instruction and student work based on Newmann's criteria.

One important aspect is the sustainability of professional development. Teachers who attended ongoing and sustained professional learning sessions to design and use authentic assessment were more effective than teachers who attended ad-hoc or short duration workshops (Koh, 2011b; Koh, et al., 2012). Teachers in Avery et al.'s (2001) study attended monthly professional learning sessions over one academic year, while teachers in Koh et al.'s (2012) study were given professional development over a two-year study period. In the latter quasi-experimental design study, teachers in the intervention group attended ongoing and sustained professional learning over two academic years. In comparison, teachers in the control group were

provided with 1-2 day workshops. Both groups involved primary 4 and 5 English, science, and mathematics teachers of average academic performance. The teachers' assessments, pre and post intervention, were rated using a rubric that was adapted from Newmann's framework. The ratings were analysed using descriptive statistics and t-tests. The five criteria in the rubric were depth of knowledge, knowledge criticism, knowledge manipulation, sustained writing, and making connections to the real world. The study reported that, following sustained and ongoing professional development, there was significant improvement in the quality of the tasks designed by teachers in the intervention group. The changes in the mean scores from the baseline period to Phase II were significantly larger for the intervention group than for the control group.

Together, these studies on teacher learning in professional development sessions built around authentic instruction indicate that the quality of teacher practice is contingent on the quality and nature of their learning opportunities. However, as this cluster of studies only involves small samples of interested teachers, their findings must be interpreted with caution and further research is required to determine if such professional development sessions yield similar results in other contexts and with other participants.

Summary of conceptual and empirical work on authentic assessment

Since Wiggins (1989) and Newmann and Archbald (1988) advanced the idea of 'authentic assessment' 20 years ago, the field has evolved significantly with scholars developing and debating the concept. The first phase of 'authentic' instruction, assessment, and achievement was largely dominated by US scholars researching on school reform. To this end, the quality of teacher assessment and student work served as indicators of reformed practices. In recent years or during what I shall refer to as *Phase II*, scholars have designed interventions to help teachers

create and develop authentic assessment, examine the definition of 'authentic' in education and analyze the impact of using authentic instruction on student achievement.

While the intervention and observational studies have reported the positive impact of authentic assessment, the findings have to be interpreted with caution. Methodologically, it is difficult to compare the overall effect of multiple large-scale studies because of the differences in design (e.g., number of raters, nature of assessments, number of schools), rating criteria (e.g., each research team devised its own rating standards), and analyses procedures. Overall, the findings from the intervention and observation studies reported the benefits to students when their teachers designed quality assessments aligned to the criteria Newmann and Associates developed. Specifically, researchers found statistically significant relationships between the quality of teacher assessment with the quality of student work. However, the findings have to be interpreted cautiously, since most intervention studies are on a small-scale, and the researchers do not report if they used comparison groups or computed the power analyses. Thus, it is difficult to determine the internal and external validity of the studies.

Substantively, comparisons were complicated because students in the studies were not taught under a common curriculum. Consequently, in addition to variations in teaching, there were also variations in what students were learning. Given that there was no common instrument to measure student performance, comparisons across studies was not possible.

Based on the review of the empirical literature, this dissertation applies Newmann and Associates' (1998) AIW standards to examine whether Singapore geography teachers' classroom assessments provide opportunities for students to demonstrate higher-order skills as envisaged by the TSLN vision. Drawing on the methodological findings from the LAAMP researchers, this

dissertation involves eight teachers, each submitting three assessments and identifying 12 students whose assignments are analyzed for this the dissertation.

Conceptualizing formative assessment

Constructivist learning theories conceive of learners as those tasked with making sense of new experiences and information by incorporating these into an existing knowledge base. Consequently, assessment practices that emanate from, and are congruent with this theory of learning, need to provide opportunities for students to build on prior learning, and apply it to novel situations. Constructivist assessment encapsulates authentic academic achievement, because both approaches value the transfer and application of skills and knowledge to new contexts, situations, and experiences. Additionally, because constructivist theories place a premium on the developmental nature of learning, formative assessment practices are a necessary and vital component of teaching and learning, especially when these practices are integrated into instruction.

This section presents the definition and history of formative assessment, the key reviews on the topic, followed by the empirical work on formative assessment from 1998 to 2011, 1998 being the year Paul Black and Dylan Wiliam (1998a) published their seminal review of the literature on classroom assessment. This publication has inspired a diversity of work on formative assessment, all with the aim of supporting and improving student learning.

Definition

Formative assessment is not new in education (Harlen, 2009). It has been defined in terms of its purpose, function, effect, timing, and usually in terms of how it is different from summative assessment. A common feature in the definitions and conceptions is that formative

assessment activities all occur within the classroom, and thus, the concept strikes a chord with classroom teachers (Cizek, 2010).

The origin of *formative* approaches in education may be traced to Socrates and his use of questions to deepen and probe understanding (Gareis, 2007), or to Lee Cronbach's work on improving course content in 1963 (Clark, 2011). However, the research community (Bennett, 2011; Cizek, 2010; Frey & Schmitt, 2007; Gardner, 2006; Roos & Hamilton, 2005; Thompson & Wiliam, 2008; Wiliam & Black, 1996) unanimously credits Australian scholar Michael Scriven for introducing the term, 'formative evaluation,' in his efforts to differentiate the formative and summative purposes of evaluation. Scriven (1967, p. 51) refers to *formative evaluation* as occurring in the "intermediate stage" in order to "discover deficiencies and successes" of any newly implemented curriculum. Based on this definition, *formative evaluation* is distinguished by its role (i.e., to identify strengths and weaknesses), timing (i.e., intermediate) and purpose (i.e., to ascertain whether the criteria used provide sufficient analysis of the goals of the curriculum program) (Scriven, 1967).

Benjamin Bloom and colleagues were the first to apply formative assessment to the context of student learning (Bennett, 2011; Newton, 2007) and mastery learning (Frey & Schmitt, 2007). It is important to note that since the work of Bloom et al., the scholarly discourse has continued to use formative assessment in relation to student learning in the classroom rather than to the evaluation of educational programs (Black & Wiliam, 2003; Clark, 2011). In their definition, Bloom and colleagues (1971) wrote of the benefits of formative assessment for students, teachers and curriculum makers

> Formative evaluation is … the use of systematic evaluation in the process of curriculum construction, teaching, and learning for the purpose of improving any of these three processes. Since formative evaluation takes place during the formation stage, every effort should be made to use it to improve the process (Bloom, et al., 1971, p. 117).

Bloom et al.'s definition of formative assessment comprises three characteristics that distinguish formative from summative assessment: purpose (formative assessment supports the learner while summative assessment is for certification and grading), timing (formative assessment occurs more frequently while summative assessment tends to take place at the end of teaching and learning), and level of generalization (formative assessment targets specific aspects of proficiency while summative tests assess broad areas of learning) (Newton, 2007).

Chronologically, in the 1980s, Royce Sadler from Australia further developed the concept of *formative assessment*. Sadler is seen as the first to propose a model of formative assessment (Shepard, 2006), and this theorization is presented in the following section, *Theorizing formative assessment*. He advances feedback as a key feature in formative assessment. However, feedback is only effective if the person or persons receiving it are able to make changes or take appropriate actions (Sadler, 1989). This concept of feedback resonates with an earlier definition by Ramaprasad (1983) who defines "feedback [as] information about the gap between the actual level and the reference level of a system parameter which is used to alter the gap in some way" (p. 4). More simply, *where the learner is going*, *where the learner is right now*, and *how to get there* (Hattie & Timperley, 2007). Feedback in the classroom serves two audiences: teachers utilize feedback to make curricular and instructional decisions, while students apply it to act on their strengths and weaknesses (Sadler, 1989).

The difference between formative and summative assessment lies in the "purpose and effect," and not the "timing" (Sadler, 1989, p. 120). Formative assessment plays a role in shaping and improving student performance as compared to summative assessment which is a collection of examples of a student's achievement status (Bloom, et al., 1971; Sadler, 1989). More recently, conceptions for formative assessment recognize that it is defined by the purposes

to which assessment information is used, rather than being perceived as a collection of procedures and instruments (Andrade, 2010).

The next phase in the conceptual history of formative assessment arose from the UK, in particular from the work of Paul Black (1998), and Wynne Harlen and Mary James (1997). Even up until today, the research on formative assessment is dominated by UK scholars. Like Scriven, and later Bloom et al., Harlen and James (1997) define formative assessment by distinguishing it from summative assessment, and in particular, in terms of validity and reliability. They suggest that summative assessment prioritizes reliability while formative assessment focuses on validity. Despite the differences, Harlen and James (1997) contend that it is "impossible in practice" and "wasteful" to ignore the fact that information gathered by teachers for formative purposes may be used to make summative judgments (p. 373). This may be achieved by ensuring that conditions are in place to guarantee that the reliability criteria required by summative assessments are met.

The current popularity of formative assessment in education is largely due to the work of Paul Black and Dylan Wiliam (1998a) whose seminal research synthesis reported that the effective use of classroom assessment would result in improvements in student achievements of between 0.4 and 0.7 standard deviations. These are significant gains because they surpass the effect sizes obtained by most educational interventions. These gains are particularly noteworthy among low achievers who made larger gains than other students (Black & Wiliam, 1998c).

This influential piece of work found its way across the Atlantic, when an easy-to-read summary of the findings, together with suggested strategies and policy implications, was published in *Phi Delta Kappan*, a popular American journal for school administrators, principals, and teachers (Brookhart, 2004). In the United Kingdom, the publication of another readable

pamphlet, *Inside the Blackbox*, was the impetus to stimulating interest in formative assessment (Brookhart, 2004).

In their review, Black and Wiliam (1998a, pp. 7-8) defined formative assessment as encompassing "all those activities undertaken by teachers, and/or their students, which provide information to be used as feedback to modify the teaching and learning activities in which they are engaged." This definition is similar to that used by Cowie and Bell (1999) in New Zealand, and Hodgen and Marshall (2005). From this definition, formative assessment involves two actions (Black & Wiliam, 1998a): learners must be aware of a gap between their current level of competence and the desired goal, and they must take action to close that gap. The idea of helping students to move from their current learning status to a desired learning goal resonates with the *zone of proximal development* (Vygotsky, 1978) and cognitive constructivism, and is also consistent with the purposes and strategies of feedback that enhances learning (Hattie & Timperley, 2007; Ramaprasad, 1983; Sadler, 1989).

In recent years, some scholars have made attempts to differentiate between 'formative assessment' and 'assessment for learning' (AfL). Formative assessment is as problematic because the term itself could be interpreted in different ways (Assessment Reform Group, 1999). The Assessment Reform Group (ARG) differentiated AfL from formative assessment: AfL is "the process of seeking and interpreting evidence for use by learners and their teachers to identify where the learners are in their learning, where they need to go to and how best to get there" (Assessment Reform Group, 2002). This definition is significant as it positions AfL as supporting both teachers and students. Teachers use formative assessment information to make instructional decisions, such as where more practice or exposition is required. More importantly, AfL also enables students to take action so as to "make progress" (Assessment Reform Group,

1999, p. 7). In this way, the ARG's conception resonates with Sadler's notion of feedback, such that teachers and learners must make purposeful use of the assessment information to effect change.

AfL clarifies the misunderstandings arising from the term 'formative' (Stobart, 2008) because in the American discourse formative assessment is often associated with continuous summative assessments that teachers use in the classroom (Bell & Cowie, 2001), and serves monitoring purposes rather than the purpose of informing instruction (Stobart, 2008). In comparison, within the British discourse, AfL and formative assessment are embedded within teaching and learning, premised on the belief that every student can achieve. They involve students engaging in self-assessment, and in turn, engage teachers and students in reviewing and making sense of assessment data (Assessment Reform Group, 1999). These essential features of AfL echo key tenets of Shepard's (2000, 2001) "emergent reform" paradigm of teaching and learning, as well as constructivist assessment theories.

Since AfL came into use in the late 1980s and 1990s , it has been seen as a "newer concept" (Gardner, 2006, p. 2) than formative assessment. AfL is believed to have been popularized in the UK by the Assessment Reform Group and in the USA by Richard Stiggins (Wiliam, 2010). The distinctive feature of AfL is that it serves to promote students' learning and should be embedded in teaching and learning. AfL only becomes formative when the information and data collected is used to shape instruction to meet students' learning needs (Assessment Reform Group, 1999; Black, Harrison, Lee, Marshall, & Wiliam, 2003a). In fact, AfL emphasizes student learning and requires teachers to use different approaches to close the learning gap, i.e., the difference between the intended goals and where students currently are at.

However, since teachers could also use information from summative assessment to adapt instruction (Thompson & Wiliam, 2008), using assessment information was insufficient to differentiate formative assessment from summative assessment. Since formative assessment is integral to teaching and learning, the "big idea" behind formative assessment is that "pupils and teachers use evidence of learning to adapt teaching and learning to meet immediate learning needs minute-to-minute and day-to-day (Thompson & Wiliam, 2008, p. 6). This revised definition integrates formative assessment closely with classroom activities and provides the clarion call for teachers to closely plan the use of formative assessment within their teaching activities. In efforts to further distill the definition, understanding, and use of formative assessment, Black and Wiliam (2009) recently re-stated formative assessment as classroom practices in which

> evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers, to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken in the absence of the evidence that was elicited (Black & Wiliam, 2009, p. 9).

This definition is significant because it advances the student's role in learning, since learners also make decisions about the next steps. This means that students are more active in learning, a conception that resonates closely with constructivist learning theories and Shepard's (2000) "emergent reform" paradigm.

A summary of the evolution of the definition of formative assessment, adapted from Brookhart (2007) is presented in Table 2.8. Since Scriven's conceptualization in 1967, formative assessment has been defined by the way in which information that is gathered about the learning process is used (Column 1). In the 1970s (Column 2) and 1980s (Column 3), in addition to providing information about the learning process, formative assessment also served the purpose of helping teachers make instructional decisions (Bloom, et al., 1971) and guiding students to

improve their learning (Column 3) (Sadler, 1989). Finally, formative assessment is also defined by its outcome, which is to motivate students (Column 4) (Black & Wiliam, 1998a; Crooks, 1988; Natriello, 1987).

Table 2.8
Evolving concepts in the definition of formative assessment[a]

(1)	(2)	(3)	(4)
Information about the learning process (Scriven, 1967)			
Information about the learning process (Bloom et al., 1971)	Purpose: For **teachers** to use for instructional decisions		
Information about the learning process (Sadler, 1983, 1989)	Purpose: For **teachers** to use for instructional decisions	Purpose: For **students** to use to improve their own performance	
Information about the learning process (Black & Wiliam, 1998a, b; Crooks, 1988, Natriello, 1987)	Purpose: For **teachers** to use for instructional decisions	Purpose: For **students** to use to improve their own performance	Outcome: To motivate students

[a]Adapted from Brookhart (2007).

As a conceptual term, *formative assessment* has been debated based on "multiple discrete purposes within a small number of misleading categories" (Newton, 2007, p. 149). Earlier conceptions focus on the differences **between** formative and summative assessment. Today, we see what can be called an 'identity crisis' **within** formative assessment. This polemical discourse is particularly vibrant in the USA, where the 'split,' as referenced by EdWeek, is between those (e.g. testing companies like Pearson) who see formative assessment as an 'instrument' and those (e.g., R. Good, 2011; Popham, 2008) who conceive of it as a "process," and as part of teaching and learning (Bennett, 2011). Broadly, each of these perspectives is an oversimplification: as an instrument, formative assessment is not effective if the process relating to its use is flawed. Conversely, formative assessment as a process is not likely to succeed if the instrument is problematic (Bennett, 2011). To this end, formative assessment is best considered a "thoughtful

integration of process and purposefully designed methodology of instrumentation" and more work is needed to develop a "strong conceptualization" (Bennett, 2011, p. 7).

Based on the above discussion, it is evident that there are benefits to be reaped from the use of formative assessment. Yet, its use in classrooms is still limited (Marsh, 2007). Some possible reasons for this dearth of practice include the privileging of high-stakes examinations, the continued public recognition of grades and achievement scores, the tendency for teachers to mimic tried and tested approaches to assessment as well as their discomfort in adopting more student-centered approaches, and finally, the pressure on principals to produce test scores that meet legislative targets (Marsh, 2007).

To conclude, two decades of work to create and delineate the identities of formative and summative assessment do not seem to have been successful (Newton, 2007). This is because the two terms reside in "qualitatively different categories" (Newton, 2007, p. 156): summative assessment refers to a specific type of assessment judgment, while formative assessment is about the "type of use to which assessment judgments are put" (Newton, 2007, p. 156). As such, attempts to create categories to differentiate the two terms can be seen as erroneous (Bennett, 2011; Newton, 2007).

This dissertation adopts the earlier Black and Wiliam definition and subsequent re-statements of formative assessment, and also use formative assessment and AfL interchangeably, a convention used by Cowie (2005), E. Hargreaves (2005), Harlen (2009), and James and Pedder (2006).

Theorizing formative assessment

This section presents the theoretical work on formative assessment from the period after the Black and Wiliam (1998a) review until the current day. However, since Sadler (1989) is

widely acknowledged as the first scholar to present a model of formative assessments (see Shepard, 2006), I have included this piece in the review. Broadly, the theoretical work on formative assessment is presented in two main categories, based on theory and empirical work.

Drawing on learning theory. Formative assessment theories by Sadler (1989, 1998), Black and Wiliam (2006, 2009), Wiliam (2010) and Shepard (2006) draw on constructivist learning theories in order to articulate general principles, frameworks, and practices. The application of constructivist theories is significant because in these formative assessment models, the student's role in the learning process is given greater prominence as scholars envisage learners as active participants, making sense of the information culled from the formative assessment activities. In this way, the theorization of formative assessment is closely aligned with the Assessment Reform Group's definition of formative assessment, because this group of academics envisaged both students and teachers as making sense of the assessment information.

Sadler's (1989) model of formative assessment is likely the earliest to appear in the field. His model combines formative assessment, feedback and self-monitoring as key elements for improving learning, and is pertinent to a diverse range of learning outcomes in different subjects. The general steps for conducting formative assessment within Sadler's model are (1) acquainting students with the goals or standards, (2) making multi-criterion judgments, and (3) involving students in self-assessment. This indicates a close association between Sadler's model of formative assessment and constructivist learning theories, which envisage students actively participating in and taking ownership of learning.

Shepard's (2006) formative assessment model is compatible with cognitive and sociocultural learning theory, and draws on the work of Black and Wiliam (1998a), Sadler (1989), and Atkin, Black, and Coffey (2001). This model serves as a tool for teachers, helping

them support student learning. It also provides information that teachers might use to improve their classroom practice (Cizek, 2010). Specifically, Shepard's model has the following key features: (1) instructional and assessment tasks reflecting learning goals, (2) learning progressions, (3) prior knowledge assessment, (4) explicit criteria and use of rubrics, (5) feedback, (6) opportunities for transfer and application of learning, (7) student engagement in self-assessment, and (8) reflection and inquiry into teaching. These principles are not intended to be used sequentially but recursively as part of ongoing improvement and revision to teaching and student learning (Shepard, 2006).

Black and Wiliam (2006, 2009) and Wiliam (2010) have devoted their efforts towards developing a theory of formative assessment for the purpose of defining and delimiting this concept within broader pedagogical theories. Their framework is intended to unify the diverse set of practices referred to in the literature as formative assessment (Black & Wiliam, 2009). This theory of formative assessment draws largely on theories of pedagogy, situated learning, activity theory, as well as models of self-regulated learning, and classroom discourse. Because of the strong pedagogical focus, Black and Wiliam's theory is related to the three key instructional processes proposed by the Assessment Reform Group: (1) determining learners' current level of learning, (2) identifying where learners are going, and (3) deciding on approaches to help learners achieve the desired goal (Wiliam, 2010). Applying activity theory, Black and Wiliam (2006) conceived four components for their theory: teachers, students and the subject discipline; teacher's role and the regulation of learning; feedback and the student-teacher interaction; and the student's role in learning. As the theory aims to unite the plethora of strategies and practices, Wiliam (2010), Wiliam and Thompson (2008), and Black and Wiliam (2012b) further contend that formative assessment could be conceptualized as comprising five

key strategies. These five strategies are a more detailed explication of those Black and Wiliam (1998a, 1998b) proposed a decade and a half ago. Specifically, the strategies are

- clarifying, sharing, and understanding learning intentions and criteria for success:
- engineering effective classroom discussions, questions, and tasks that elicit evidence of learning
- providing feedback that moves learners forward
- activating students as instructional resources for one another; and
- activating students as the owners of their own learning (Wiliam, 2010, p. 31)

The significant contribution of this theory is that the strategies are not simply a list of successful approaches. Rather, in developing a formative assessment theory, the above five strategies serve as "lenses" to view more closely and to think about practice, especially in the areas of psychology, pedagogy and curriculum (Wiliam, 2010, p. 37). The value of this theorizing effort is the integration of other theories to formative assessment, thereby creating a comprehensive and holistic approach that encapsulates definitive features of this concept.

Drawing on empirical work. Two theoretical frameworks describe the types of formative assessment practices teachers use in their classrooms. These two frameworks were constructed based on empirical findings of research with young children. The frameworks are significant because they describe and define the ways teachers approach and practice formative assessment. Broadly, both frameworks distill a continuum of formative assessment practices based on teachers' interactions with students, and decision-making during classroom instruction. Both frameworks draw on constructivist learning theories, and in particular, sociocultural theories.

One formative assessment framework posits that teachers use two main types of formative assessment: planned and interactive (Cowie & Bell, 1999). This framework was based on the findings of a two-year empirical research project with 10 elementary teachers in New Zealand. *Planned* formative is used deliberately as part of whole class instruction and the

strategies used include gathering, interpreting and acting on the assessment information. This type of formative assessment practice is narrower, more specific, and concentrates on gathering information to inform teaching and instructional planning. Conversely, *interactive* formative assessment strategies such as noticing, recognizing and responding arise spontaneously during interactions with students to address their learning needs in groups or individually. This approach to formative assessment is more spontaneous, and typically occurs when teachers and students interact. Successful use of both types of formative assessment requires teachers to be skilled in their 'pedagogical content knowledge' (Shulman, 1987) in order to make judgments needed to support student learning, and to guide instruction (Cowie & Bell, 1999). While both types of formative assessment appear to lie on diametrically opposing poles, Cowie and Bell (Cowie & Bell, 1999) do not advocate one type over the other. Instead, they recommend a blended and iterative use of planned and interactive formative assessment practices.

Pre-emptive formative assessment, an extension of planned formative assessment, denotes teacher responses taken to remedy students' understanding before misconceptions arise to make learning ineffective (Carless, 2007). Specifically, Carless (2007) conceptualizes pre-emptive formative assessment to create a further distinction between individual and whole-class assessment, and argues that creating this awareness would provide teachers with a larger repertoire of strategies. Drawing on constructivist learning principles, pre-emptive formative assessment recognizes the importance of feedback strategies. Feedback that occurs following the completion of a task or activity tends to be ineffective because most students do not have the opportunity to respond to or effect changes upon receiving the feedback. To remedy this gap in practice, Carless (2007) postulates that the close daily interactions between classroom teachers and their students provides a ready context for teachers to clarify understandings before the

misconceptions lead to ineffective learning or performance, or with a given task, the loss of points or marks. As such, pre-emptive formative assessment is "anticipatory feedback" that supports student learning (Carless, 2007).

To effect pre-emptive formative assessment, Carless proposes that teachers decide on the mode (whether planned or interactive), target group (work with students as individuals or as a class), and timing (pre-emptive or reactive). There are some limitations to this theory. First, learners may be overly dependent on teachers to guide them before they work on the task. Next, the specific focus on certain misconceptions may lead to teaching-to-the test, which raises the issue of whether improved performance during high stakes examinations is equated to learning. Third, weak and highly variable teacher assessment literacy may aggravate learning misconceptions, thereby requiring professional development. Finally, teachers need to make ethical judgments to justify the types of advice and practices before acting on them.

The second formative assessment framework is informed by learning theories, and authentic and formative assessment. This conceptualization of formative assessment practices is based on classroom interaction and linguistic structures (Torrance & Pryor, 1998). It positions formative assessment as "intersubjective social processes situated in, and accomplished by, interaction between students and teachers" (Torrance & Pryor, 1998, p. 616).

The model adopts a normative approach as it is characterized by two "ideal-typical" (p. 616) dimensions: *convergent* and *divergent*. Both dimensions are conceived to be a continuum of possible approaches that teachers use to support learning, and as such, emerge from teachers' views of learning. However, these distinctions are more "heuristic than descriptive"; they are tendencies and each is not exclusive of the other (Torrance & Pryor, 1998, p. 153). *Convergent* formative assessment is underpinned by behaviorist learning theories and is used by teachers for

summative purposes directed towards specific learning goals (Torrance & Pryor, 2001). Conversely, *divergent* formative assessment leverages on social constructivist learning theories; thus, it focuses on students' understanding, with the purpose of discovering what students understand, know or can do (Torrance & Pryor, 2001). Divergent formative assessment embraces a view of learning where students collaborate to develop new ideas and in the process, delegates more autonomy to the learner (Miller & Lavin, 2007).

Continuing with Torrance and Pryor's work and drawing on empirical research conducted in infant classrooms through post-graduate education over an eleven-year span, Pryor and Crossouard (2008) propose a socio-cultural theorization of formative assessment. This is based on the observation of teachers and learners as they interact and respond to student work during formative assessment. Thus they position formative assessment as a "discursive social practice, involving dialectical, sometimes conflictual processes" (Pryor & Crossouard, 2008, p. 1). Their theorization is drawn from multiple theoretical traditions, including sociocultural learning theories, Vygotsky and the Cultural Historical school, and Bernstein's concept of framing and classification. Broadly, it recognizes that, when teachers and learners interact, there are issues of power at play, especially when they negotiate and make meaning of the formative assessment tasks and criteria. Furthermore, this social nature of formative assessment necessitates that teachers and learners re-construct and re-shape their identities as they interact to understand the task and the associated criteria. This shift of identities links back to the divergent-convergent formative assessment continuum. Within convergent assessment, the teacher is an assessor and a teacher, while in divergent assessment, the teacher is both an educator and a learner. Pryor and Crossouard (2008, p. 16) also introduce "metacontextual

reflection" to represent an element in which the clarification and deconstruction of criteria, and issues of power and control take place.

Summary. All in all, the inductive and deductive theorization efforts point to attempts to unify a somewhat disparate field. This is especially seen in the theorization of formative assessment advanced by Black and Wiliam (2006, 2009) in which they seek to unify practices, theory, and teacher-student roles. To this end, theorization efforts resonate with the attempts to define formative assessment. For instance, as formative assessment is perceived as being integral to teaching and learning, the theories identify or provide pedagogical approaches that are useful in formative assessing (Black & Wiliam, 2006; Shepard, 2006). Common themes across the theories include involving students in self-assessment, enabling them to participate more actively in the classroom, and giving them ownership of learning (Black & Wiliam, 2006; Sadler, 1989; Shepard, 2006). The theories also seek to describe and understand the nature of formative assessment practices (Cowie & Bell, 1999; Torrance & Pryor, 1998) based on teachers' interactions with students as individuals and as a class (Carless, 2007). The latter is a significant development because the theorization identifies the issues of power and control in constructivist classrooms, which require teachers and students to re-negotiate and re-shape their roles and identities in the classroom when applying formative assessment strategies.

Existing literature reviews on classroom assessment

This section presents a summary of the review of the literature on classroom assessment. The term used here is 'classroom assessment' because the reviews focus on assessments conducted in the classroom, not external, high-stakes assessment. Second, the reviews concentrate on both formative and summative classroom assessment. Here, I use classroom assessment because not all the studies included in the reviews focus on formative assessment.

However, this body of work on teacher assessment in the classroom is significant for a number of reasons.

First, it provides the pioneering research on, and foundation for understanding the current interest in formative assessment. The key pieces by Crooks (1988), Natriello (1987), and Black and Wiliam (1998a) provide a history of the evolution of this field, highlight the benefits and challenges of classroom assessment, and identify the value and limitations of the empirical work in this area. Although this dissertation is interested in examining the state of the field *after* Black and Wiliam's (1998a) seminal work, I include the reviews by Natriello and Crooks to present a comprehensive picture of the evolution of the field, and also because these two pieces are widely cited, especially in the way they use and define classroom assessment. It is worth noting that Crooks and Natriello used 'evaluation' in their reviews, and a decade later, the word was replaced by 'assessment.' Second, although Black and Wiliam made an initial attempt to define formative assessment, current scholars continue to refer to and examine these pieces in their attempts to deepen, define, and delineate the concept of formative assessment. Third, the way scholars thematize and present their findings from these reviews provides an indication of the key areas of research interest in classroom assessment. I will discuss the commonalities of the three reviews before briefly presenting a fourth review by Brookhart (2007) which updates the field after Black and Wiliam. As the seminal work in this field arose from Black and Wiliam's work, their review is discussed the greatest detail.

The reviews by Crooks, Natriello, and Black and Wiliam are similar in that they all studied the impact of classroom evaluation practices on learning strategies, motivation, and achievement. However, their approaches diverge in that they examined different aspects of evaluation: Natriello applied the widest definition of evaluation, including certification, selection,

direction, and motivation, while Crooks and Black and Wiliam limited their reviews to classroom assessment. All three reviews also differed in terms of the number of studies reviewed. Natriello examined 91 studies while Crooks reviewed 241 studies. Black and Wiliam examined 250 pieces of empirical work to explore whether changes in classroom assessment improve learning. It is significant that just nine studies were common to the Crooks and Natriello reviews (Black & Wiliam, 1998a, 2003), suggesting a disparate field at the time and the complexities involved in defining formative assessment (Black & Wiliam, 2003).

An analysis of the way the researchers thematize and present their findings provides an indication of the key areas of research interest in classroom assessment. Natriello (1987) focused on eight stages of the evaluation process, namely: establishing the purposes, assigning tasks, setting criteria, setting standards, sampling information, appraising, providing feedback, and monitoring. In comparison, Crooks (1988) focused on the impact and use of normal classroom testing practices, other instructional practices, and the motivational aspects relating to classroom evaluation. Finally, Black and Wiliam presented their findings on the impact of classroom assessment in five categories: teachers, students, strategies and tactics, systems, and feedback.

All three reviews find that classroom evaluation affects students in many ways and is "one of the most potent forces influencing education" (Crooks, 1988, p. 467), to the extent that there is an overemphasis on summative evaluation, and too little emphasis on assisting students in learning (Crooks, 1988). Black and Wiliam provide the most explicit estimate of the impact of classroom assessment – they announce that the effect size of using classroom assessment ranges from 0.4 to 0.7 standard deviations. These gains are substantial and significant for two reasons: they are larger than the usual effect sizes in educational research, and larger gains are

found in low achieving students as compared to the other students. Contextualizing these statistics for a lay audience, the Assessment Reform Group (1999) explains that the magnitude of the reported effect translates to be the equivalent of between one or two grades at the General Certificate for Secondary Education (GCSE) for one student. Alternatively, using TIMSS as an international benchmark, this means a leap for England – moving from the middle position of 41 countries to the top five (Assessment Reform Group, 1999).

Despite the benefits of formative assessment, Black and Wiliam (1998c) were concerned that teachers' assessment practices leaned towards tasks aligned to behaviorist theories (e.g., rote learning rather than higher-order thinking), and that these practices emphasized superficial learning by placing more emphasis on the quantity and presentation of work than on the quality of work. In addition, when teachers privilege the grading and ranking of students, students have lower self-esteem, become de-motivated, and are demoralized. The undue emphasis on grades and summative functions leads teachers to shift their attention from providing advice for improvement and learning to pursuing marks. Another problematic practice is feedback, which, when provided, serves managerial and administrative purposes rather than showing students how to learn effectively, or how to advance their learning.

To effect the types of learning gains reported in the research, Black and Wiliam (1998a) positioned formative assessment as a critical leverage point in classroom practice - it is the "heart of pedagogy" (p.16). There are five formative assessment strategies: (1) provide effective feedback; (2) involve students in the learning process; (3) use the assessment information to shape instruction; (4) be aware of the impact of assessment on students' motivation and self-esteem; and (5) involve students in self-assessment and guide them to address their learning gaps (Assessment Reform Group, 1999). Feedback is a strategy also recommended by Crooks and

Natriello. However, for its use to be effective, it has to be specific, relevant to the area of need, and be provided in a timely manner so that it is relevant (Crooks, 1988). Finally, implied in the recommended strategies is the need to link formative assessment with new conceptions of learning (Shepard, 2000, 2001).

Despite the positive effects on student achievement, Black and Wiliam (1998a) are cautious about the rigor of their findings as many of the studies they reviewed lacked ecological validity; thus some of the strategies may not be easily replicated. In addition, they only provided a range for the effect size, rather than an exact estimated value because for this area of study, the "assessments vary greatly in their sensitivity to instruction" (Wiliam, 2010, p. 21). Furthermore, underlying differences in many studies, despite the common focus on learning gains, make any combination of their results less meaningful (Black & Wiliam, 1998a).

Most recently, Brookhart (2007) has advocated an "inclusive view" of summative and formative assessment, arguing that both have a role in the way "assessment works for learning" (Brookhart, 2007, p. 45). Her review categorized practices in nine categories: conventional assessment practices, teacher beliefs, classroom assessment environment, effective formative assessment and feedback, negative effects of poor classroom assessment, student motivation, student involvement in assessment, validity and reliability, and effects of formative classroom assessment on student achievement. Brookhart's findings are similar to those of earlier reviews – that formative assessment provides students with the information to improve, and the confidence to do so. More importantly, formative assessment, when properly used, does not conflict with external, summative assessments (Brookhart, 2007).

Empirical research on formative assessment

This section presents the empirical work relating to formative assessment from 1998 to 2011. This date parameter was selected to begin with the review of the empirical work based on Black and Wiliam's (1998a) review of the literature.

This review of the empirical work on formative assessment is guided by the 'systematic review' process (Evans & Benefield, 2001) which requires designing an explicit research question, being transparent in the review methodology, using precise criteria for including or excluding studies, and providing clear conclusions drawn from the review (Hammersley, 2001). To ensure rigor, only peer-reviewed articles that applied classroom assessment, from primary through pre-university education, in the spirit of Black and Wiliam were selected (see Black, et al., 2003b; Black, Harrison, Lee, Marshall, & Wiliam, 2004; Black & Wiliam, 1998a, 2009). In total, this review of the literature on formative assessment includes 54 studies of which six pieces of research are presented in more than one category. This review includes scholarly contributions from 12 education jurisdictions [Barbados, Canada, Hong Kong, Ireland, Italy, the Netherlands, New Zealand, Nigeria, Singapore, Spain, United Kingdom, and USA]. Probably due to the links with Black and Wiliam, the largest number of studies originates in the United Kingdom. The empirical work on formative assessment is mostly dominated by interventions (39 out of 54 studies involved an intervention, either in the form of trying out a formative assessment strategy or teachers undergoing a professional development program).

Black and Wiliam (1998a) did not use a meta-analysis to report their findings because of the differences in the quality and nature of the studies reviewed, and the methodologies used. Similarly, while this review used systematic processes, I did not attempt to synthesize the conclusions because the studies provide insufficient data to compute an overall effect size, and

many studies used qualitative methods. Furthermore, the empirical work featured is better understood by presenting common patterns and critiquing differences (Tierney, 2006).

The review of the empirical work on formative assessment is presented in four categories; namely *Pedagogue* (an old, formal English word for teacher); *Pupil*; *Procedures and tools;* and *Policies*. The studies are categorized based on their dominant research questions. Coincidentally, these categories are largely aligned to the ones Black and Wiliam used in 1998.

Background

Since Black and Wiliam's (1998a) influential review, the field of formative assessment has been replete with studies examining how different strategies and approaches enhance its use, computing the effect of its use, and understanding how teachers and students make sense of this instructional concept. The impressive effect sizes that they reported have spurred interest in the use of formative assessment in teaching and learning. Perhaps that is why of 54 studies included in this review, more than half involve an intervention.

While there are many formative assessment interventions, the majority of studies do not adopt randomized control experiments. Since these intervention studies do not employ the random assignment of teachers to students, it is difficult to estimate causal relationships from the studies. Many of the studies included in this review are small-scale mixed methods (n=21) or qualitative research (n=32) methods studies. Within the studies included in this review, just five research pieces involved the use of quantitative methods. The predominance of mixed methods and qualitative research methods is a salient point, suggesting that researchers are using multiple data sources and analyses methods to examine this concept. I suggest that the use of mixed methods and qualitative research illustrates the importance of understanding and uncovering learning processes, thoughts, and perceptions—a view that is consistent with the philosophy of

formative assessment, especially with the new ways of theorizing formative assessment (see Black & Wiliam, 2006, which were presented under 'Conceptualizing formative assessment'; 2009; Crossouard, 2009; Wiliam, 2010).

Formative assessment and the Pedagogue

In understanding educational change and implementation, House's (1978, 1981) technological, cultural, and political perspectives are germane to the review of the empirical research on formative assessment. The technological perspective views innovations as systematic and precise; the cultural perspective considers beliefs, value systems, and shared meanings; and the political perspective embodies negotiation, tension, and resolution (House, 1981). Applying the cultural perspective (House, 1978, 1981) helps in understanding how and in what ways teachers interpret and integrate innovation like formative assessment into their beliefs and value systems (A. Hargreaves, et al., 2002). Teachers achieve this by reconciling themselves with the purpose of formative assessment. From the cultural perspective, teachers employing formative assessment want to understand its philosophy, and align it with their beliefs and value systems before blending it into their practice. House's cultural perspective (1978, 1981) suggests that researchers should be mindful of how the change is communicated and understand how teachers perceive their roles in a formative assessment classroom. The presentation of the research in this category is divided into three groups, namely, perspectives, teachers' practices, and decisions shaping formative assessment practices.

Pedagogue's perceptions of formative assessment. The empirical studies presented in this section focus on developing understanding, and transformation among teachers who are the agents of formative assessment change in the classroom. The research examined teachers' conceptions of assessment and learning (Brown, 2004; Brown, Kennedy, Fok, Chan, & Yu, 2009;

Colby-Kelly & Turner, 2007; E. Hargreaves, 2005; Remesal, 2007, 2011; Segers & Tillema, 2011), and their roles and responsibilities in formative assessment (Gioka, 2009) within the context of second language assessment, accountability, and national curriculum policy. Understanding teachers' conceptions and perceptions of assessment is important because these influence how they teach (Brown, 2004).

The studies in this cluster use a mixture of qualitative, quantitative, and mixed methods approaches. As these are observational studies, the empirical research examines teachers' conceptions and understanding of formative assessment using lesson observations (Gioka, 2009), interviews (Gioka, 2009; Remesal, 2011), and surveys (Brown, 2004; Brown, et al., 2009; E. Hargreaves, 2005; Segers & Tillema, 2011). Lesson observations, interviews, and the collection of artifacts provided rich descriptions of the activities in the classroom (Gioka, 2009). A novel survey approach using qualitative research methods involved respondents writing their responses anonymously to the prompt 'assessment for learning' (E. Hargreaves, 2005). The responses of the 83 teachers were then analyzed using a grounded theory procedure which involved the grouping and re-grouping of the comments based on three main themes – assessment, learning, and assessment for learning. The findings are presented later in this section.

Another type of survey study used to capture teachers' perceptions of assessment involved the use of a close-ended survey. The Teachers Conception of Assessment (TCOA) survey was first developed and validated in New Zealand, before being adapted and used in the Netherlands (Segers & Tillema, 2011) and Hong Kong (Brown, et al., 2009). The original survey instrument used in New Zealand was comprised of 65 items and intended to be a self-report attitude inventory (Brown, 2004). The survey contains four categories of teachers' conceptions of assessment: improvement, school accountability, student accountability, and irrelevance (of

assessment). Teachers responded to the survey based on a 6-point scale (Brown, 2004). Participants' responses (n=525) to the New Zealand survey were analyzed using structural equation modeling, factor analysis and Multiple Analysis of Variance. This same survey instrument was subsequently adapted for use in a study to examine Hong Kong teachers' (n=288) conceptions of assessment following the introduction of an assessment reform in Hong Kong (Brown, et al., 2009). The Hong Kong reform aimed to increase the emphasis on assessment *for* learning and to reduce the pressure on assessment *of* learning. The survey adapted for Hong Kong involved a translation and a shorter rating scale (it was a 4-point rating scale compared to the 6-points used in New Zealand). Statistically, a shorter rating scale enabled researchers to examine if there were a socially-acceptable bias in teacher responses. Brown's survey was also adapted by a team of researchers from the Netherlands (Segers & Tillema, 2011). This team used an abridged version of the survey which had 27 items, and analyzed the responses of 351 Dutch teachers using Maximum Likelihood analyses and multiple factors analyses.

One pattern in the findings indicated that teachers conceive of classroom assessment in two broad categories: assessment for learning/formative assessment and assessment of learning/summative assessment (E. Hargreaves, 2005) or assessment associated with monitoring teaching and learning and assessment related to certification and accountability (Remesal, 2011). This is especially the case when there are curriculum policies that encourage formative classroom practices in addition to high stakes summative assessments. Where this tension occurs, the weight of teachers' responses indicate that they lean heavily towards the "assessment as measurement" and "learning as attaining objectives" categories (E. Hargreaves, 2005).

A second pattern in the findings indicates differences in the conceptions of assessment between primary and secondary teachers. When compared to their secondary colleagues,

Spanish primary teachers focused on assessment practices that are pedagogically-inclined while their secondary colleagues privileged the accrediting conceptions of assessment (Remesal, 2007, 2011). These findings point to the need to incorporate the differences between primary and secondary teachers' practices if significant changes are to take place in Spain (Remesal, 2007).

Third, the patterns of teachers' conceptions of the functions of assessment differ by country. In relation to the Teachers Conception of Assessment survey, there were similarities and differences in perceptions of assessment among the New Zealand (Brown, 2004), Hong Kong (Brown, et al., 2009), and Dutch (Segers & Tillema, 2011) teachers who responded to Brown's (2004) survey. In terms of the four functions, teachers from New Zealand, Hong Kong and the Netherlands agreed with the *improvement* function of assessment. Unlike their New Zealand counterparts, Dutch teachers did not conceive of assessment as related to *school accountability*. While the Hong Kong teacher's conception of assessment included *student accountability*, this perception was not shared by the New Zealand and Dutch teachers.

The empirical work sought to explain teachers' formative assessment practices. One reason for the type of formative assessment practices teachers adopted is the apparent contradiction in educational policy (Gioka, 2009; E. Hargreaves, 2005). For example, policy may encourage teachers to increase formative assessment practices in the classroom, yet continue with the high-stakes use of summative assessment outcomes. The prevalence of the English National Curriculum and national assessments continued to push teachers towards the measurement model of assessment and learning (E. Hargreaves, 2005). This impacted the way teachers saw their roles in the classroom – they were either examiners or teachers or both (Gioka, 2009). For instance, science coursework is a component in the UK's General Certificate of Secondary Education (GCSE) and Advanced Level examinations, and the policy intent was for

this component to be taught as a 'process.' However, because the component contributed towards students' performance in the national examinations, the two science teachers participating in Gioka's (2009) study did not appreciate "process" objective but viewed it as a "final product" to which they assigned a grade (p. 424). As a result, Gioka (2009) recommended that science teachers working on the coursework component make efforts to shift their conceptions of the task, refocusing on learning instead of emphasizing the grade. However, caution is required when applying the conclusions from this study because it only reports the practices and perceptions of two teachers.

Pedagogue's formative assessment practices (Observational studies). This section presents the observation studies (n=7) in which researchers documented the range of teachers' formative assessment practices (Gioka, 2006; I. Lee, 2007; Rea-Dickins & Gardner, 2000; Riggan & Oláh, 2011), the reasons for the use (Bell & Cowie, 2001), the way teachers use formative assessment (Gattullo, 2000; Pryor & Torrance, 1998; Riggan & Oláh, 2011), especially within a policy context (Kirkup, 2006; Volante & Beckett, 2011). Some of the studies examined the use of formative assessment within a specific discipline (Bell & Cowie, 2001; Hodgen & Marshall, 2005), while others looked at general practices (B. Marshall & Drummond, 2006).

Scholars examined the nature of the classrooms in which formative assessment was being introduced (Webb & Jones, 2009), as well as the challenges faced by early-adopters of formative assessment (Carless, 2005). These studies primarily focused on teachers' formative assessment practices in core subjects in the curriculum such as mathematics (n=3), science (n=2), and English (n=5), which are assessed in national and state tests; there were no studies on the humanities, the aesthetics, or physical education. A common thread running through the findings

is the tension that teachers face when balancing the use of formative assessment within the context of statutory national testing.

The observational studies documenting teachers' formative assessment practices involved lesson observations (Gattullo, 2000; Gioka, 2006; Pryor & Torrance, 1998; Riggan & Oláh, 2011), interviews (Dibu-Ojerinde, 2005; Gattullo, 2000; Gioka, 2006; Pryor & Torrance, 1998; Riggan & Oláh, 2011; Volante & Beckett, 2011), analysis of student work and other artifacts (Dibu-Ojerinde, 2005; Gattullo, 2000; Gioka, 2006; Riggan & Oláh, 2011), questionnaire surveys (Dibu-Ojerinde, 2005; Kirkup, 2006), case study (Kirkup, 2006), and focus group discussions (Kirkup, 2006). The use of mixed and multiple research methods presents a rich description of the diversity and quality of teachers' practices, with the use of different data sources to triangulate the findings.

On the whole, teachers use a variety of formative assessment practices; many of these strategies were recommended by Black and Wiliam (1998b) in *Inside the Black Box*. These strategies include feedback (e.g., Dibu-Ojerinde, 2005; Kirkup, 2006), peer assessment (e.g., Riggan & Oláh, 2011), student self-assessment (e.g., Kirkup, 2006; Volante & Beckett, 2011), and questioning (e.g., Volante & Beckett, 2011). Despite the use of this plethora of strategies, the overall conclusion is that practices were weak (Gioka, 2006) or were merely directed towards improving achievement, rather than promoting learning (Dibu-Ojerinde, 2005; Gattullo, 2000; Kirkup, 2006; I. Lee, 2007; Riggan & Oláh, 2011). One example of weak formative assessment practice is that teachers did not communicate assessment criteria to students, a key tenet of formative assessment practice (Gioka, 2009). Another evidence of weak formative assessment practice is that teachers focused on administrative tasks like correcting mistakes instead of developing meta-cognitive skills (Gattullo, 2000).

Teachers' formative assessment practices involved their interpreting data (e.g., from written work), and then adjusting their instructional plans during the re-teaching period. The most popular approach for this was feedback. However, more often than not, feedback and instruction were geared so specifically and precisely that students were not given the latitude to incorporate the assessment information. These approaches indicate that teachers' formative assessment practices "converge" (Torrance & Pryor, 1998) towards the narrowing of learning because students are directed to the correct response. For instance, one study examining the practices of 32 mathematics teachers in the USA reported that they used formative assessment in an organizational way, in particular to identify weak content areas or students within a class, rather than to address student misconceptions (Riggan & Oláh, 2011). Specifically, these teachers implemented interim assessments, interpreted the data, and then adjusted their instruction accordingly during the re-teaching period. While such assessment practices are formative, in that teachers incorporate the information into and adapt their teaching, Torrance (2007) has expressed concern that such approaches change the practice from assessment *for* learning to assessment *as* learning, because the procedures and strategies completely dominate the learning experience.[3] A similar pattern of teachers' feedback practices "converging" towards student achievement was reported in a Nigerian study (Dibu-Ojerinde, 2005). This study found that the 300 teachers teaching in private schools in five Nigerian school districts rarely provided students with formative feedback or adapted their instruction based on assessment information. An examination of students' notebooks indicated that on occasions when teachers did provide feedback, the purpose converged towards student achievement, rather than to shape learning.

[3] Torrance (2007) uses 'assessment as learning' differently from Canadian scholar Lorna Earl who uses the phrase to argue for a congruence between learning and assessment practices aligned with student self-assessment. Her use, according to Torrance, is consistent with the spirit of Assessment for Learning. Torrance's use of 'assessment as learning' arises from a concern with the "displacement" of learning by "procedural compliance" and what he calls "achievement without understanding" (p.293) and is particularly germane to post-secondary training qualifications.

However, as the study was confined to private secondary schools, these conclusions may not be representative of all Nigerian teachers.

The nature of teachers' formative assessment practices is influenced by educational policy and reform. Yet, in a context of high-stakes testing, teachers from different countries respond differently. In the United Kingdom, teachers struggled to find a balance between different assessment purposes, and they reconciled these tensions by privileging summative assessment (Kirkup, 2006), or by devising methods to make formative use of summative assessment data, a strategy recommended in Black, Harrison, Lee, Marshall and Wiliam's (2003b) *Assessment for Learning: Putting it into Practice*. Findings from a survey of 490 head teachers and primary school teachers across the United Kingdom indicated that the respondents were able to integrate the assessment data from national assessment in the classroom to support teaching and learning (Kirkup, 2006). The data showed that while teachers used feedback and student self-assessment, the aims were to improve achievement, rather than to promote learning. Despite schools making a concerted effort to integrate summative and formative approaches, teachers faced challenges using these strategies (Kirkup, 2006). This finding is contextualized within the climate of mandatory national testing and the use of league tables to rank schools in the UK.

Compared to the reactions from teachers in the United Kingdom, there were different responses from Ontario teachers in their reception to the *Growing Success: Assessment, Evaluation and Reporting in Ontario Schools* policy which aimed to increase teachers' assessment literacy. Broadly, Ontario teachers used a diverse range of formative assessment strategies (Volante & Beckett, 2011), including the use of questioning, the provision of feedback instead of grades, peer and self-assessment, and the formative use of summative assessment;

once again, this list resonates with the recommendations by Black and Wiliam (1998c) and Black et al. (2003b). Despite this diversity of practice, interviews with 20 teachers indicated that there was an imbalance in the use of specific types of formative assessment strategies, and in particular, those that were associated with improvements in student learning (Volante & Beckett, 2011). At the same time, teachers reported that they were starting to value learning over the undue emphasis on grades. Finally, while teachers viewed large-scale assessments negatively, nearly all teachers in this sample from this south-central province analyzed the EQAO results (the provincial tests) to some degree.

The responses of the teachers from the United Kingdom and Ontario indicate that education innovations such as formative assessment strategies challenge teachers' conceptions and practices. As such, teachers need time to interpret and integrate the change into their values and beliefs (A. Hargreaves, et al., 2002; House, 1978, 1981). Making sense of teachers' reactions to and uses of formative assessment might be more complex than simply learning various theories since the former requires a wider corpus of knowledge than learning theories (Pryor & Torrance, 1998). One way of understanding how teachers and students interact during formative assessment is to apply psychological and sociological theories when examining such interactions during routine assessment "events" (Pryor & Torrance, 1998, p. 151). This enhances the understanding of the realities of the classroom, and is a more purposeful way to understand the nature of formative assessment interactions between teachers and students (Pryor & Torrance, 1998).

Pedagogue's formative assessment practices (Interventional studies). The studies (n=6) presented in this section also focus on the *what* and *how* of teachers' formative assessment practices, the difference being that the researchers introduced an intervention to the study.

Typically, the intervention involved teachers attending a professional development program (e.g., Hodgen & Marshall, 2005), specially commissioned research project (e.g., Bell & Cowie, 2001; B. Marshall & Drummond, 2006) or a curricular remediation course (e.g., Rea-Dickins & Gardner, 2000). This cluster of studies offers a way to compare the nature and quality of teachers' formative assessment strategies after they had attended some professional development sessions, as compared to the studies above which were purely observation and survey studies.

One theme running through these six studies is the range and nature of formative assessment practices that teachers use, and the challenges and problems they face in doing so. These studies recognize that there are generic formative assessment strategies, such as those suggested by Black and Wiliam. At the same time, differences in each subject require teachers to employ approaches consistent with the respective discipline (Bell & Cowie, 2001; Hodgen & Marshall, 2005; Rea-Dickins & Gardner, 2000). For instance, a study of formative assessment practices among ten primary and secondary science teachers in New Zealand identified overarching characteristics: (1) is responsive, (2) based on information and evidence, (3) is a tacit process, (4) requires professional knowledge and experience, (5) is an integral part of teaching and learning, (6) is carried out by both teachers and students, (7) has specific purposes, (8) is contextualized, (9) creates dilemmas, and (10) involves student disclosure (Bell & Cowie, 2001). These practices were distilled from a study over a two-year period. However, eight of these formative assessment practices were largely circumstantial, and so, to use formative assessment, teachers had to apply professional knowledge and experience in responding to students, recognize and integrate formative assessment within teaching and learning, and work closely with students in the assessment process (Bell & Cowie, 2001).

Subject-specific formative assessment approaches are appropriate because the structure of knowledge differs between subjects (Black & Wiliam, 2012b; Hodgen & Marshall, 2005). Hodgen and B. Marshall's (2005) study in the UK examined the use of formative assessment strategies in an English lesson and a mathematics lesson. Based on in-depth qualitative analyses of the lessons, the findings indicated that on the surface, the purposes and strategies appeared to be similar (Hodgen & Marshall, 2005), and even resonated with the generic pedagogical techniques developed by Black et al. (2003b). However, the successful use of formative assessment in different disciplines needs to be anchored in the specific content and pedagogy of each subject (Hodgen & Marshall, 2005). This was especially so because subjects like English and mathematics are contrasting disciplines and they approach knowledge from different perspectives (Hodgen & Marshall, 2005). This finding points to the significance of helping teachers develop 'pedagogical content knowledge' (Shulman, 1987). At the same time, teachers should be careful of "proceduralising" formative assessment but rather draw on the good practices of different subjects (Hodgen & Marshall, 2005, p. 172).

The second theme emerging from this cluster of studies focuses on understanding, improving, and supporting teachers' use of formative assessment after attending professional learning sessions. As formative assessment resonates strongly with constructivist learning theories, the quality of teachers' use of such strategies requires them to embrace new cultures of teaching and learning (B. Marshall & Drummond, 2006), to negotiate and understand the required changes in procedures, culture, and roles in the classroom (Webb & Jones, 2009), and to work with and manage colleagues, administrators and parents who might be resistant to these new methods of teaching (Carless, 2005). The different social contexts affect the way teachers construct their roles in formative assessment classrooms. At the classroom level, several studies

point to "reculturaling" (A. Hargreaves, 1994) the contexts to support the use of formative assessment. More specifically, changes in teachers' practices are associated with changes in classroom culture (Webb & Jones, 2009). Under activity theory, using formative assessment strategies or "mediating artifacts" (Webb & Jones, 2009, p. 174) require open, supportive and trusting classrooms so that students are comfortable with speaking up, and are unperturbed about their peers' perceptions of them (Cowie, 2005).

Furthermore, both teachers and students need to review their beliefs about the nature of learning, and re-conceive their roles and agency in the classroom. Professional development introduces teachers to formative assessment and builds their capacity to use it. However, teachers whose lessons most closely mirror the spirit of assessment for learning are more likely to value and provide learner autonomy as their overall goal in teaching (B. Marshall & Drummond, 2006). Once teachers develop such an attitude, they will not be fazed by obstacles such as school constraints, student ability or national assessments (B. Marshall & Drummond, 2006).

Finally, the quality of teachers' formative assessment practices is contingent on the complex interplay of personal, micro, and macro forces—the three components from Clarke and Hollingsworth's model of professional development. These forces were found to be significant in Carless' (2005) study of two novice Hong Kong teachers' experiences implementing formative assessment while engaging in action research cycles. The findings indicated that there were changes in teachers' attitudes towards the formative assessment approaches as the project evolved, especially when this was matched with support and positive feedback by Carless, the lead researcher. Building on the Clarke and Hollingsworth's original model, Carless' exploratory model reflects the complexity of change at the personal, micro, and macro levels. The personal domain pertains to teacher knowledge and beliefs (e.g., the vanguard of teachers' convictions);

the micro level relates to school support (e.g., the lack of collegial and school support, presence of researcher support), and the macro level points to the external environment (i.e., the similarities and differences between AfL and the earlier policy). Given this complexity, professional development alone is insufficient to bring about change. Rather, within this context, change would take time, especially since the baggage from the previous policy was still in place.

Decisions shaping formative assessment practices. While the studies in the preceding segment explored the *what* and *how* of teachers' formative assessment practices, the cluster of studies included in this section (n=3) focuses on the *how* and *why* of teachers' practices. Understanding *how* and *why* teachers enacted formative assessment practices provides an insight into the theory-practice gap, illustrating that while teachers recognize the significant value and contribution of formative assessment to student learning, their ability to carry out these strategies in the classroom is complex, problematic, and contingent on larger social, political, and contextual factors, of which contradictory messages from national curriculum and assessment documents constitute one key factor.

Teachers' formative assessment practices are guided by curricula policies, which may lead to a gap between their beliefs and practices. This was illustrated in a study of 12 geography teachers from the United Kingdom. Interviews with these teachers indicated that three criteria influenced their design and use of assessment tasks: conceptualization of formative assessment, the National Curriculum level descriptors, and professional craft knowledge (Tiknaz & Sutton, 2006). However, the evidence from lesson observations showed that the feedback and assessment practices of these teachers were narrow and closely aligned to the national curriculum descriptors. This indicated that teachers' formative assessment practices were

heavily influenced by the prevailing objectives-driven view of learning as indicated by the descriptors mapped out in the National Curriculum for geography.

National assessment policies also influence teachers' assessment practices. A study of 558 teachers' self-report responses to a questionnaire following a series of formative assessment workshops identified principles shaping teachers' formative assessment practices: teachers valued "making learning explicit" and "promoting learning autonomy" but disagreed with the "performance orientation" of assessment (James & Pedder, 2006, p. 130). The first two practices are associated with assessment *for* learning whilst the third is tied to assessment *of* learning. The evidence from this study suggests that teachers are striving to find that balance between the "performance orientation" dimension and the other two factors (James & Pedder, 2006). Arguably, in the UK's severe assessment climate at the time of the study, teachers had little choice but to reconcile their beliefs in a classroom culture "informed by pedagogic values" and that developed by policymakers (James & Pedder, 2006, p. 131). The implications of the study for teacher-educators and school administrators are the need to support teachers in narrowing the value-practice gap for greater fidelity to the formative assessment rhetoric.

In addition to external influences, teachers' assessment practices are shaped by personal factors. Their ability to use new practices following professional development is contingent on having time and space to integrate and adapt new forms of assessment into their existing repertoire. In the Classroom Assessment Project to Improve Teaching and Learning (CAPITAL), Coffey, Sato and Thiebault (2005) used a case study approach to examine how two middle school science teachers adapted their classroom practices as they incorporated formative assessment strategies in their professional practice. Throughout the four-year research project, teacher-participants met researchers to discuss their emerging practices and reflect on their

practices. Lessons observed were presented as case study narratives to highlight the different approaches used by the teachers. The findings indicated that responding to the change was a highly personal experience for teachers, as they adapted and integrated this form of assessment into their existing practice, usually guided by their beliefs and theoretical knowledge (Coffey, et al., 2005). Decisions that teachers made are driven by their underlying beliefs about teaching, what is valued in the educational process, how teachers view themselves as professionals and as persons, and who their students are. This study illustrates that classroom change is highly contextualized and personal for teachers, and one way to realize change is to provide teachers with sufficient time to internalize and integrate the ideas within their practices (Coffey, et al., 2005).

Summary. Several patterns relating to teachers' use of formative assessment emerged from this cluster of studies. Broadly, teachers hold both formative and summative conceptions of assessment. However, the weight of their responses towards one or the other depend on the existing pressure exerted by the national assessment and accountability systems (E. Hargreaves, 2005). Second, primary level teachers' conceptions of assessment are closely associated with their pedagogical orientations while secondary teachers' perceptions of assessment are more linked to assessment for accountability purposes (Remesal, 2011). Third, there are variations across countries in teachers' conceptions of assessment in three areas, namely that assessment provides information for improvement, and that it is used for both school and student accountability (Brown, 2004; Brown, et al., 2009; Segers & Tillema, 2011).

In terms of teachers' practices, the empirical work reveals that teachers use a variety of formative assessment strategies (e.g., Kirkup, 2006; Volante & Beckett, 2011), most of which are recommended by Black and Wiliam (1998c). Overall, the research findings report that

teachers' formative assessment practices are weak – a similar conclusion to Black and Wiliam (1998a). Teachers practices are either so precise that they result in the narrowing of learning (e.g., Riggan & Oláh, 2011) or they do not adhere to the key tenets of formative assessment strategies, such as communicating assessment criteria to students (Gioka, 2006). Additionally, while there are generic formative assessment strategies, disciplinary differences require the use of subject-specific approaches (Bell & Cowie, 2001; Hodgen & Marshall, 2005; Rea-Dickins & Gardner, 2000).

Teachers' use of formative assessment is shaped by their personal beliefs (Coffey, et al., 2005), and educational policy (E. Hargreaves, 2005; Tiknaz & Sutton, 2006). Because formative assessment is underpinned by theories like constructivist learning theories, teachers need time to understand, incorporate and integrate the spirit and intent of such assessment into their beliefs and value systems. They also need to establish new classroom cultures and re-perceive the role of themselves, and their students in the classroom (Webb & Jones, 2009).

The strength of the studies presented in this section lies in the use of multiple data sources which triangulate to provide rich information about the activities in the classroom and the views of teachers. They used a variety of research methods – quantitative (e.g., Brown, 2004; James & Pedder, 2006), qualitative (e.g., E. Hargreaves, 2005), and mixed methods (e.g., Dibu-Ojerinde, 2005). Additionally, as the empirical work is drawn from several countries, similarities in findings across different countries suggest that the patterns are not unique to a particular country.

However, some of the findings need to be interpreted cautiously because of the research designs. In several of the studies, researchers did not explain the procedures used to recruit participants. Those that did provide details, generally examined formative assessment practices

of participants who were already converts to formative assessment. This suggests that there exists a positive bias, such that these teachers' views are not representative of their colleagues. Additionally, the sample of participants used in many of the studies was small (e.g., two teachers in the study by Hodgen & Marshall, 2005). While these issues are not new – Natriello (1987), and Black and Wiliam (1998a) previously expressed concern about the research designs in their studies – they serve to remind readers to be critical and discerning when surveying the field.

Formative assessment and the Pupil

Students are often on the receiving end of educational research studies. However, their voices are frequently omitted from the research discourse, despite many scholars' (e.g., Levin, 1994; Levin, 2000; Rudduck, 1991, 2002; Rudduck & Flutter, 2004) calling for their inclusion. Eliciting student voices during educational change is a way of acknowledging that they are active agents in the learning process (Kirton, Hallam, Peffers, Robertson, & Stobart, 2007). Of the 54 empirical studies included in this review, only four pieces had research questions that were interested in capturing students' comments or thoughts about formative assessment. Yet, students are the targets of all sorts of assessment, formal and informal, formative and summative.

Understanding students' views guides researchers and educators to develop instructional strategies that are more closely aligned to the learners' needs. Just as teachers take time to grow into their roles in a formative assessment classroom, students also need help to take on the new roles of active learners within the formative assessment classroom. The studies in this section are presented in two categories; the first examines students' perceptions and views about formative assessment, and the second analyzes the impact of formative assessment on student learning.

Pupils' perceptions of formative assessment. Constructivist learning theories envision teachers' roles shifting from experts to facilitators, and students' roles changing from passive to

active participants as they are accorded greater ownership of their learning. In "divergent" (Torrance & Pryor, 1998, 2001) assessment practices, students have more autonomy, and are actively collaborating with their peers in constructing knowledge. To this end, students who have been socialized to "convergent" (Torrance & Pryor, 1998, 2001) assessment practices need support to reconstruct their roles in a formative assessment classroom. When the classroom culture or environment lacks trust and respect, students may be less inclined to participate in class activities (Cowie, 2005), and consequently, may not appreciate or understand the benefits of formative assessment practices (Colby-Kelly & Turner, 2007).

Students' reactions to formative assessment are complex (Colby-Kelly & Turner, 2007; Cowie, 2005). Overall, their views are positive, but may be contradictory (Cowie, 2005). This conclusion is based on a study in New Zealand that examined the experiences of 114 grades 7-10 students who attended a specific science lesson. On the one hand, students conceived themselves as active learners, and indicated that they preferred teachers' oral feedback to be provided as suggestions, rather than as "closed" comments, so that they could actively make sense of the comments (Cowie, 2005). Yet, when it came to completing work, students reported that they favored specific feedback so that they could be directed to complete the task (Cowie, 2005).

Students are sensitive to the way feedback is provided (Colby-Kelly & Turner, 2007; Cowie, 2005). The Grades 7-10 students from the New Zealand study maintained that they wanted feedback to be given individually or within a small peer group – they were afraid of being perceived as inadequate or failing to catch up. Therefore students needed to feel that there was sufficient trust and respect between teachers and their classmates before asking questions or clarifying their understanding (Cowie, 2005). Students also wanted their teachers to exercise discretion when proving feedback (Colby-Kelly & Turner, 2007).

There are several reasons why students respond in this manner. One conjecture is that they are socialized to a particular learning discourse, and so, some formative assessment strategies like peer review may discomfort them. Consequently, students need to re-shape the way they participate in the classroom. From the New Zealand study, it is inferred that a disconnect between teaching and learning is possible if teachers and students have different expectations and understandings of the purposes of and approaches to formative assessment. These findings suggest that students require time to clarify their roles and expectations within the formative assessment classroom.

Students' backgrounds and characteristics could be another possible reason for their responses to formative assessment practices. Older, high-achieving students tend to have positive perceptions of both formative and summative assessment. This was found in a study of 50 students from Grades 10-12 taking honors classes in the USA (Brookhart, 2001). Through interviews, surveys, and lesson observations, the perceptions of these high-achieving students fell into two categories: self-assessment, and the integration of formative and summative assessment (Brookhart, 2001). These students were able to relate their performance in classroom assessment to the standards that they needed to achieve, to engage in regular self-assessment, and to actively use assessment information to improve their learning. As grades mattered to these students, they viewed summative assessment as a means of providing them with signposts along their education journey. These students did not distinguish formative and summative assessment, but successfully integrated both types. While this study illustrates the way high achieving students utilize assessment information, these findings need to be interpreted with caution because these students are by many yardsticks, "privileged" (Brookhart, 2001, p. 159).

The effective use of formative assessment practices requires students and teachers to hold similar conceptions of assessment. However, there is often a gap between teachers' and students' conceptions, as shown in a study involving 712 students and 351 teachers in the Netherlands (Segers & Tillema, 2011). In this study, both teachers and students completed the Conception of Assessment surveys developed in New Zealand by Brown (2004).[4] The responses indicated that teachers and students saw assessment as having a school accountability purpose. However, the two groups differed in the way they viewed the purposes of formative and summative assessment. While students differentiated between assessment that supported learning and assessment for student accountability, teachers did not distinguish between these two aspects. The dissonance between teacher and student responses is an area for further research because this disconnect affects the effective use of AfL strategies in a way that benefits students and their learning (Segers & Tillema, 2011).

Impact of formative assessment on student outcome. The intervention studies presented in this section report the impact of student learning on achievement in specially-designed assessments (e.g., embedded assessments in Yin et al., 2008) or existing assessments. The latter may be school-based (e.g., Wiliam, Lee, Harrison, & Black, 2004, used end-of-module science tests) or external (E. Smith & Gorard, 2005) assessments. With the exception of two studies (i.e., E. Smith & Gorard, 2005; Yin, et al., 2008), the empirical work reports that using formative assessment strategies has a positive impact on student outcomes. This conclusion is consistent with Black and Wiliam's (1998a) findings.

There are gains or improvements to student learning based on the use of formative assessment. Of the nine studies presented in this section, seven reported gains or improvements. The largest effect size was from the Kings Medway Oxford Formative Assessment Program

[4] This study was reviewed under *Formative assessment and the Pedagogue*.

(KMOFAP) led by Black and Wiliam and colleagues (see, Black, et al., 2003b, 2004). This intervention was novel in that it did not prescribe any particular formative assessment protocol or tool, but encouraged teachers to develop their own action plans. In total, the 24 participating teachers developed 102 activities, averaging four plans per teacher. To ensure rigor in the comparison, the project team set up the best possible comparison group while not disrupting the program of the school. This involved using a parallel class taught by the same teacher the previous year or by a different teacher. The researchers computed statistical means to standardize the differences between the experimental and comparison groups. Analyses of test scores indicated that students taught by KMOFAP teachers had higher gains compared to peers in the comparison group (Wiliam et al. 2004). The mean effect size due to the innovation was 0.32. However, this finding must be viewed cautiously because the schools and participating Local Education Authorities (LEA) were not randomly selected. The project team had approached these schools because their LEA was already known to be supporting formative assessment work. The researchers also acknowledged that the comparisons between the control and comparison classes were not "equally robust" (Wiliam, et al., 2004, p. 62) because of the research design. For example, in some schools, the same teacher taught the control and the intervention classes whilst in another school different teachers taught the control and the intervention classes.

Other than changes in test scores, student learning was evidenced by qualitative changes in student outcomes (Cooper & Cowie, 2010; Davies, Durbin, Clarke, & Dale, 2004), as indicated by deeper understanding (Davies, et al., 2004), application of knowledge to real world contexts (Fox-Turnbull, 2006), student decision-making (McDonald & Boud, 2003) and self-esteem (Miller & Lavin, 2007). In New Zealand, when teachers had the autonomy to choose a

formative assessment strategy pertinent to their learning goals, the outcome was that students were more focused on learning, and developed deeper understanding of an idea (Cooper & Cowie, 2010).

Similar qualitative improvements to student work are reported in a study that focused on student self-assessment and the use of explicit assessment criteria in geography in the UK. In this study, students were given a set of quality criteria and asked to identify the level they wanted to aim for in their geography assessment (Davies, et al., 2004). Each time they completed their work on a unit, these students assessed their work using the criteria. Analyses of the baseline and post-test data indicated that after applying the self-assessment process, students in the intervention group received, on average, higher than predicted scores in the year end geography assessment. In the second year, when students were asked to select their best piece of work from a collection of tasks, and then to explain their choice, students in the intervention group discussed the 'quality' of the work while those in the control group spoke about the 'quantity' of information in the piece of work. The strength of this study is that it involved one control and one intervention class in each of three schools. However, while these findings are encouraging, they are interpreted cautiously because the participating teachers were interested in carrying out research scholarship, and the students involved in the study were from the upper half of the ability range in each class.

Using formative assessment strategies such as feedback enables students to apply and transfer classroom learning to real world contexts. Primary level students in the control group (n=17) in a New Zealand study worked on an 'out-of-context' task and those in the experiment group (n=36) worked on an 'out-of-context' followed by an 'in-context' task. The 'in-context task' was designed as an authentic technological practice and was completed after content

instruction and teacher feedback. From content analysis of students' work, Fox-Turnbull (2006) reported that the experiment group produced detailed and feasible work-solutions after the 'in-context task.' These students were also more confident in explaining and justifying their decisions. For teachers to provide accurate and effective feedback, procedural, conceptual, societal and technical knowledge – aspects of pedagogical content knowledge (Shulman, 1987) – are needed.

The effective use of formative assessment has a positive impact on students' self-esteem (Miller & Lavin, 2007), and decision making skills (e.g., students in the study by McDonald & Boud, 2003, said that they were better able to choose careers suited to their personality). When pre-university students in Barbados were taught self-assessment strategies, the intervention resulted in significant gains among the students who applied these strategies. The reported effect sizes were 0.26, 0.13 and 0.24 for the humanities, science and business courses respectively (McDonald & Boud, 2003). Students also reported that learning to apply self-assessment skills in daily decision-marking helped them become more analytical, independent, and empowered (McDonald & Boud, 2003). The use of self-assessment strategies had more impact on the self-esteem of low ability students than on that of their higher-ability peers (Miller & Lavin, 2007). This was reported in Miller and Lavin's (2007) study which examined the changes in self-esteem of 370 upper primary students in Scotland. The 16 teachers who taught these students used a range of formative assessment strategies as part of the pilot implementation of Scotland's *Assessment is for Learning* initiative. The use of the Rosenberg self-esteem survey both before and after the assessment indicated higher gains among the low ability students, especially those with low self-perceptions.

Formative assessment strategies may not always have a positive impact on student learning. Complex interacting factors hinder the effects of formative assessment strategies on student achievement and other outcomes. To examine the impact of formative assessment on student achievement and motivation, one study in the USA provided 12 science teachers in the experiment group with embedded formative assessment[5] tasks and trained them to use formative assessment strategies to teach a unit of science (Yin, et al., 2008). Student outcomes were measured using a motivational questionnaire and several achievement tests, including multiple-choice items and performance assessment. The analysis using Hierarchical Linear Models revealed that the assessments embedded in the curriculum did not have a statistically significant effect on students' motivation, achievement and conceptual change. Possible reasons for this finding are the difficulty of conducting an experiment, the teachers' (in)effective use of the strategies and the lack of timely feedback given to teachers (Yin, et al., 2008).

Another reason why formative assessment strategies may not be associated with positive gains in student learning might reside within students themselves. Students are socialized to a particular way of schooling, and interventions that disrupt this equilibrium make them disconcerted and frustrated, or result in their resisting any upheavals to customized practices (Black, et al., 2003b). Typically, such students are socialized to focus on getting right answers (Black, et al., 2003b). Welsh students studying in a comprehensive school became upset when their teachers applied a strategy of writing comments instead of giving them grades and marks (E. Smith & Gorard, 2005). Not receiving marks and grades was upsetting for the students because they used these to direct their efforts at improvement. These secondary students had mostly negative reactions to the use of feedback in the experiment, possibly because the purpose of the

[5] Details of the embedded formative assessments are not reported in this research paper but in another paper by Shavelson et al. (2008) and Ayala e al. (2008). In this paper, the researchers only feature the instruments and results of the study.

experiment had not been explained to them. After a year, using linear regression analyses, the researchers reported that test scores for the treatment group in English, mathematics, and Welsh were lower than that of the control group. In science, the difference was not statistically significant. The findings remind researchers that if interventions are designed to transfer some control of learning from teachers to students, then the full cooperation of and understanding by students are needed.

Summary. Students' voices have not been included in much of the empirical work on formative assessment, yet researchers are active in devising interventions to examine the impact of formative assessment on student learning. The studies in this section that examine student perceptions of assessment indicate that they have distinct preferences in the way they value and apply formative assessment. This is especially so for older and high achieving students (Brookhart, 2001). A significant finding indicates that when using formative assessment, especially feedback, teachers must be sensitive to students' feelings (e.g., Colby-Kelly & Turner, 2007).

Overall, the empirical studies on student outcome report the positive impact of formative assessment on student achievement. Following the exposure to one or more formative assessment strategies, seven studies reported higher student achievement scores (e.g., Black, et al., 2003b; McDonald & Boud, 2003) and other outcomes such as deeper understanding (e.g., Davies, et al., 2004). In one study, the impact of formative assessment strategies on motivation or achievement was not statistically significant (Yin, et al., 2008). Finally, in one study, students expressed their displeasure over the use of experimental formative assessment strategies (E. Smith & Gorard, 2005).

Once again, caution is needed in the interpretation of the findings in this cluster. Some of the studies lack external validity because the teachers involved in the study are a self-selected group, as they, their schools or their local educational authorities were already converts to the formative assessment movement. For example in the KMOFAP study, the schools and participating LEA were selected because the leading educators understood and supported the study (Black, et al., 2003b). Furthermore, the studies are mostly small-scale in nature, with a few exceptions like the study by MacDonald and Boud (2003) which sampled over 100 students. Thus, the findings need to be viewed critically before making claims of the positive contribution of formative assessment to student achievement.

Formative assessment and Procedures and Tools

The benefits of formative assessment have prompted scholars to investigate different ways to improve student learning. The thirteen studies included under *Procedures and Tools* present work that examines the effectiveness of strategies (e.g., Ruiz-Primo & Furtak, 2007), and the effectiveness of fidelity to a strategy (e.g., Furtak et al., 2008). The strategies investigated in the studies were those recommended by Black and Wiliam (e.g., Parr & Timperley, 2010, used feedback), or those specially designed by researchers (e.g., Ruiz-Primo & Furtak, 2006b). In terms of the effectiveness of strategies, the empirical work examined student learning with respect to achievement scores (e.g., Furtak, et al., 2008) and student comments (e.g., MacPhail & Halbert, 2010). In this category, *procedures* refer to strategies that have prescribed steps and stages, while *tools* involve the use of specially designed instructional manipulatives (e.g., Chin & Teou, 2009, use cartoons) or assessment tasks (e.g., 'rich tasks' in MacPhail & Halbert, 2010).

Formative assessment strategies. This section is in two parts. The first briefly presents the conceptual and empirical work on feedback, which is viewed by Black and Wiliam (1998a),

Brookhart (2007), Crooks (1988), and Natriello (1987) as a key formative assessment strategy. The second part covers disparate strategies designed and developed by different scholars. The findings indicate the variations in the use of feedback and suggest the need for greater support to enable teachers to use these strategies more effectively.

The definition of feedback dates to Ramaprasad (1983) who conceptualized it as "information about the gap between the actual level and the reference level of a system parameter which is used to alter the gap in some way" (p. 4). However, feedback is ineffective unless the learner or user has the means to use the information to close the gap (Sadler, 1989). Three conditions enable a student to use the feedback for self-monitoring, namely being cognizant of the standards, making multicriterion judgments, and having direct evaluative experience (Sadler, 1989). Close examination of these conceptions of feedback suggests that they are aligned with constructivist learning theories because they are based on "the assumption that human thought can operate at various levels" (Roos & Hamilton, 2005).

The value of feedback is widely acknowledged, and its use and effectiveness have been collated in a number of reviews, most recently by Hattie and Timperley (2007), and Shute (2008). In both reviews, the scholars build on Ramaprasad's and Sadler's models of feedback to provide a structure for effective feedback. Table 2.9 provides a summary of the approaches to operationalize Ramaprasad's definition and to bring about effective and useful feedback.

Table 2.9
Conceptions of feedback

	(1)	(2)	(3)
Ramaprasad (1983)	Availability of data on the reference level of the system parameter	Availability of data on the actual level of the system data	Availability of a mechanism for comparing the data on the reference level with that on the actual level to generate information about the gap between the two levels
Sadler (1989)	Establish a goal or reference point	Compare the actual performance level with the goal	Engage in appropriate action to close the gap
Hattie & Timperley (2007)	Where am I going?	How am I going?	to next?
Shute (2008)	Motive (student needs it)	Opportunity (student receives in time to use)	Means (student is able and willing to use)

Two recent reviews of feedback report impressive effect sizes related to its use. The values range from an average of 0.5 (Hattie & Timperley, 2007) to 0.8 or higher (Shute, 2008). Feedback is most effective when it is directly related to a task and shows students how to complete or conduct the assignment (Hattie & Timperley, 2007). Conversely, low effect sizes are computed when feedback is given as a form of extrinsic motivation, such as in the form of praise, rewards, and punishment (Hattie & Timperley, 2007). An example of a large effect size arising from the use of feedback was reported in a study from New Zealand. This study examined the quality of 56 teachers' written feedback on their students' writing assignments following two years of professional development. It was found that there was an association between the quality of teachers' written feedback and the progress in student writing (Parr & Timperley, 2010). The effect size of the intervention was 1.19. Since teachers' ability to provide quality feedback is an indicator of teacher knowledge, pedagogical content knowledge is strongly associated with the ability to provide quality feedback (Parr & Timperley, 2010).

Issues related to feedback as a formative assessment strategy include the timing of feedback (Hattie & Timperley, 2007; Shute, 2008), the nature of the feedback (Hattie & Timperley, 2007; I. Lee, 2007), the impact of feedback on student learning (Parr & Timperley, 2010), and student comments about their teachers' use of feedback (Williams, 2010). The empirical work reported that teachers' use of feedback for writing classes continues within the teacher-dominated paradigm, in which feedback 'converges' toward the provision of correct answers, while at the same time, continuing to relegate students to a passive role in the teaching and learning process. Lee (2007) explored the nature of feedback that 26 Hong Kong teachers provided in their writing classrooms. From content analysis of teachers' feedback comments which were interpreted using an analytic framework based on formative assessment categories, Lee concluded that teachers continued within a traditional teaching paradigm, in which their markings and comments directed students toward the right answers. The feedback to written work did not provide latitude for students to exercise any agency in correcting or locating errors.

Student voice in teaching and learning is important in informing educators what pedagogical practices meet students' needs. The empirical work in this cluster explores students' responses to the use of feedback (Williams, 2010), and looks at success criteria and student ownership (Read & Hurford, 2010). These studies provide evidence that even young primary level students are able to articulate and identify strategies that enhance their learning. A mixed methods study with 56 students from Year 8 in New Zealand (ages 12-13 years old) used a questionnaire to elicit students' views about the use of a list of feedback strategies (Williams, 2010). Students were asked to comment on and explain why they thought each strategy was or was not helpful. The second part of the study used 'story' items to elicit students' comments on topics like the purpose, timing, and usefulness of feedback. The responses from this small group

of students (n=2 girls, n=2 boys), indicated that on the whole, students in this sample had a good understanding of the purpose of feedback and were able to identify aspects of feedback that they found useful (Williams, 2010).

Like their older counterparts, primary students in the UK were also able to identify their own success criteria. A class of Year 5 students in one study were invited to co-construct and use a set of success criteria in the 'Continuum' (Read & Hurford, 2010). Different continua were developed for different skills (e.g., risk taking, reading, and an archaeology theme) and each required students to write criteria indicating their goals at three points in time: the present, a mid-point, and some end point. For each criterion and time point, students identified strategies (or 'struts') that would help them attain their end goal. One of the researchers taught students how to apply the strategy. The analysis of student work indicated that primary-age students were able to identify their own success criteria. The 'struts' provided teachers with insights into students' perceptions of their learning needs (Read & Hurford, 2010). However, because the strategy was only used on three occasions, the researchers called for students to be given opportunities to practice using and refining the success criteria. One weakness of the study is that during the study period, the researchers served as peripatetic teachers, and the outcome might differ if the actual teacher had conducted the pilot.

The nature and fidelity of using a formative assessment strategy is critical to its success. More specifically, the research reports that teachers whose assessment practices re closely aligned to the procedures developed by researchers have students with higher test scores. Three studies from the USA focused on examining how teachers used a four-step (Ruiz-Primo & Furtak, 2006b, 2007) assessment conversation, and the fidelity with which they applied it (Furtak, et al., 2008). These studies were premised on the assumption that the effectiveness of

the embedded assessments from the Foundational Approaches in Science Teaching (FAST) program were contingent not only on the quality of the specially-designed prompts, but also on the way they were implemented. In these studies, "assessment conversations" were used as informal interactive formative assessment (Ruiz-Primo & Furtak, 2006b, p. 207). This formative assessment procedure for science had four steps: teacher Elicits, Student responses, teacher Recognizes, and teacher Uses (ESRU). The researchers recorded and coded the lessons to compare the way in which four teachers enacted the ESRU procedure. Using students' pre- and post-intervention test scores, a general linear model analyses was conducted to estimate if students' change in scores differed by teacher. The findings revealed that on average, teachers who used the four-step procedure more closely achieved higher student scores. Overall, there was a positive impact on students' achievement scores. While the researchers recognized that the sample was small, they rationalized that this enabled them to study the assessment conversations in detail.

Formative assessment tools. In addition to examining Black and Wiliam's (Black & Wiliam, 1998b, 1998c) five strategies, scholars have devised tools to enhance the use of the strategies (e.g., Chin & Teou, 2009, used concept cartoons for peer and self-assessment) or developed new tools (e.g., MacPhail & Halbert, 2010, used 'rich tasks'). The aim of the studies was to introduce new tools to widen the teachers' pedagogical repertoire (e.g., Leat & Nichols, 2000) or to refine a planning framework with assessment instructions (e.g., MacPhail & Halbert, 2010).

The impact of these tools was analyzed with reference to students' conceptual understanding (e.g., Aschbacher & Alonzo, 2006), student engagement (e.g., Crossouard, 2011), and teachers' own learning (e.g., Feldman & Capobianco, 2008). In terms of eliciting student

understanding, the analysis of students' written work in tools like notebooks (e.g., Aschbacher & Alonzo, 2006) and concept cartoons (e.g., Chin & Teou, 2009) enabled teachers and researchers to identify gaps in learning or misconceptions. One study examined whether student work in science notebooks served as a useful means of formative assessment (Aschbacher & Alonzo, 2006). The researchers provided professional development for 25 teachers of Grades 4 and 5 in a California school district in science content knowledge, unit learning goals, assessing student work, and feedback. They analyzed students' understanding by rating their work in the notebooks, a performance task, and pre-and-post multiple-choice tests. Based on linear regression analyses, Aschbacher and Alonzo (2006) reported that the notebook scores for the teachers in the study predicted student achievement on the performance test and the post-multiple-choice test, thus concluding that the notebooks produced an accurate reflection of student knowledge. In another study, students' gaps in science knowledge and misconceptions were diagnosed and identified through the use concept cartoons (Chin & Teou, 2009). These tools were set up as part of peer and self-assessment strategies, and used to stimulate talk and argument in science in a study with two classes of primary students in Singapore (Chin & Teou, 2009). Student learning was captured through audio-recordings and through the use of drawings to articulate learners' ideas, and. The analysis of these audio conversations and drawings provided a record of students' thinking (Chin & Teou, 2009). In a third example, strips of 'mysteries' were used by 25 geography teachers in a study in the UK (Leat & Nichols, 2000). Students had to move these 'mysteries' or clues around to 'solve' the mystery. The researchers used observations and student interviews to understand the cognitive processes underpinning the physical movement of the clues (Leat & Nichols, 2000). With each subsequent session, students

became more efficient in reflecting on their reasoning as they manipulated the clues to solve the 'mysteries,' and consequently, teachers provided less scaffolding (Leat & Nichols, 2000).

The value of these tools depends on the way teachers set up the tasks (e.g., MacPhail & Halbert, 2010), use the tools (e.g., Aschbacher & Alonzo, 2006), and accept this method of assessment (Leat & Nichols, 2000). The nature of the designed tasks provides useful tools for formative assessment. In one study, the 'rich' authentic tasks that were designed resonated strongly with the authentic assessments envisaged by Newmann, Secada, and Wehlage (1995), and Wiggins (1990, 1993). This 'rich task' is a student self-assessment tool called an 'assessment wheel.' It required students to record, reflect on, and map their learning in response to the rich task and to assess their journey toward pre-set objectives (MacPhail & Halbert, 2010). Like the feedback models, the tool identifies learning gaps, and requires students to take action by planning for the next phase after self-assessment. This study is unique in that it is the sole study in this review that focuses on the development of an assessment framework for Physical Education (P.E.), a subject considered non-core in many national curricula. From interviews with teachers and from focus group sessions with students, MacPhail and Halbert (2010) reported that the quality of student learning and engagement during P.E. lessons improved after learning to use the self-assessment strategies in the "assessment wheel." Students reported that they enjoyed being given more responsibility for their learning. The strength in this study rests on the design of a sustained intervention (continued for one year). A weakness of the study is that the limitation of data did not allow the researchers to explore the association between the 'rich' assessment task and its impact on the nature and extent of learning in P.E. (MacPhail & Halbert, 2010).

The impact of the procedures and tools is contingent on the way teachers use them. In the study of science notebooks, qualitative analysis of ten sampled notebooks indicated problematic issues in the way teachers used these notebook tools (Aschbacher & Alonzo, 2006). For instance, some teachers simply tasked students to copy correct responses into the notebooks. In addition, the effective use of the information in the notebooks depends largely on the teachers' science content knowledge and their commitment to student learning vis-à-vis the need to complete the curriculum (Aschbacher & Alonzo, 2006). To harness the value of the notebooks for science learning, professional development, provided prior to and during the intervention, was beneficial in helping teachers understand the subject content, analyze the entries, and use productive feedback strategies (Aschbacher & Alonzo, 2006). Finally, teachers had to be comfortable with the use of these tools. In the 'mysteries' study, teachers had some misgivings about this somewhat informal nature of the assessment process. In spite of this, Leat and Nichols (2000) rationalized that teachers' concerns could be accommodated and the 'mysteries' strategies could be seen as a useful assessment approach.

Summary. The studies presented in *Formative assessment and Procedures and Tools* provide qualitative (e.g., Read & Hurford, 2010) and quantitative (Parr & Timperley, 2010) evidence of the value of formative assessment strategies. The research suggests that formative assessment is not just a tool or instrument per se, but involves the use of approaches such as "assessment conversations" (Ruiz-Primo & Furtak, 2006b) or feedback that are embedded in or integrated into teaching and learning. Another significance of these studies is that, in addition to focusing on core subjects (e.g., science and writing in Chin & Teou, 2009 and in Parr & Timpereley, 2010 respectively), the impact of formative assessment strategies on non-core subjects (e.g. in MacPhail & Halbert, 2010) was also examined.

The findings reported from the studies on feedback are significant in several ways. First, the effective use of feedback has positive impact on student achievement (Parr & Timperley, 2010), as evidenced by the large effect sizes. Second, the use of feedback strategies is enhanced by incorporating student voice (e.g., Read & Hurford, 2010). The empirical work provided evidence that students, even those at the primary level, are able to identify which feedback strategies support them in learning. Third, specially designed feedback procedures may improve the way teachers provide feedback, and the fidelity to the steps in the strategy is associated with student achievement (Ruiz-Primo & Furtak, 2006a).

The empirical work indicates that specially designed tools for formative assessment provide opportunities for teachers to diagnose and to identify misconceptions in student learning and understanding. However, the effective use of these tools depends on the way teachers design tasks (MacPhail & Halbert, 2010), make use of the tools (Aschbacher & Alonzo, 2006), and value the use of these tools for teaching and learning (Leat & Nichols, 2000). Finally, the reported findings need to be interpreted carefully because the research designs do not enable the inference of causal effects; the plethora of strategies, research methods and analyses make comparison between studies difficult; there is limited information provided on how participants were recruited for the study; and the studies use small samples that make generalization to other contexts difficult.

Formative assessment and Policy

The empirical work presented under *Policy* examines the changes in classroom assessment practices in relation to new educational and curriculum policy. These studies are presented at the national, school, and classroom levels. For each level, the studies illuminate how systemic structures may facilitate or impede the implementation and sustainability of formative

assessment. They also demonstrate the complexity of the change process, highlighting that beyond the technological and functional aspects, cultural and political elements come into play in the use of formative assessment. Together, the technological, cultural and political elements echo House's (1978, 1981) perspectives of educational change. Too often, attention is only devoted to the technical aspect, which is about building expertise, learning strategies, and providing change input. The cultural aspect of educational change pays attention to the meanings that teachers and students attach to formative assessment, while the political perspective presents the tensions and power issues that require skillful negotiation before implementing formative assessment. From the policy perspective, the interaction of these three perspectives is necessary for successful implementation of change (House & McQuillan, 1998), and this is exemplified by the findings reported in the empirical work presented under this category.

National policy. The sustainability of implementation of formative assessment at the national level requires concerted and coordinated efforts to ensure effective implementation and continued adoption. These structures were documented in the evaluation of Scotland's *Assessment is for Learning* (AifL) policy (Hayward, 2007; Hayward & Hedge, 2005; Hayward & Spencer, 2010; Hutchinson & Hayward, 2005; Kirton, et al., 2007) and England's *Primary Strategy: Excellence and Enjoyment* (Boyle & Charles, 2010). England's *Primary Strategy* called for alternatives to obtain "measurable excellence in teaching and learning" and for educators to use assessment information "in support of learning" (Boyle & Charles, 2010, p. 287). Scotland's AifL aimed to develop a coordinated assessment system in which assessment serves formative, summative, and accountability purposes (Hayward & Spencer, 2010). The mixed methods evaluation study in the first phase of AifL was conducted in 35 primary schools and junior high schools, and investigated the adoption of this policy following the release of the

2000 report, 'Improving Assessment in Scotland.' In England, the qualitative study of teachers' understandings of formative assessment, and changes in practices following the *Primary Strategy* involved a questionnaire, interviews, and lesson observations in 43 schools, and was conducted six years after implementation.

There were differences in the implementation of the policies in both jurisdictions. In Scotland, the findings indicate that there were changes in teaching and learning practices, in that students became more active in their learning and more engaged in peer and self-assessment. The landscape was diverse, with a myriad of practices being used for different subjects and levels in the schools (Kirton, et al., 2007). During interviews, teachers identified three clusters of ideas which they attributed to the sustained engagement in AifL practices: educational integrity, personal and professional integrity, and systemic integrity (Hayward & Spencer, 2010). One weakness in practice which highlighted the complexity of sustainable change was the depth of teachers' understanding of learning and teaching under AifL. Teachers only perceived formative assessment as a collection of strategies which they religiously used. They lacked the "language of the … theories" (Hayward & Spencer, 2010, p. 171) and consequently, did not engage deeply with the spirit and intent of formative assessment. Ultimately, fidelity of implementation to the AifL intent was a complex endeavor with "no simple solution," since all schools were taking different pathways along a "common journey" (Hayward & Spencer, 2010, pp. 172-173).

Comparatively, in England, there was little evidence of change as practices still focused on summative assessment (Boyle & Charles, 2010). The findings pointed to schools having a "blatant misunderstanding" of formative assessment and its purposes (Boyle & Charles, 2010, p.

298). Formative assessment practices were weak because pedagogy focused on coverage and pace, rather than on depth (Boyle & Charles, 2010).

There are several reasons for the differences in the implementation of these two assessment policies. In both jurisdictions, careful attention was devoted to providing technical assistance. In Scotland, there was a well-executed systemic support structure provided at the local and regional levels to assist schools (Kirton, et al., 2007), while in England, there was government funding and the supply of consultants (Boyle & Charles, 2010).

The genesis of AifL drew heavily on the KMOFAP model, with Scottish policymakers taking feedback from teachers seriously, especially in their request for incremental change. Like Black, Wiliam and colleagues' KMOFAP, the Scottish change process was a grassroots affair – each school developing and designing its own assessment action plan after professional development sessions (Kirton, et al., 2007). There was evidence of changed cultures in that communities of practice became distinctive platforms to support teachers in their endeavors to use and apply formative assessment strategies. This implementation approach indicates respect for teachers' professional knowledge and a recognition that schools would respond with different programs fitting their specific diverse student population. Comparatively, in England, Boyle and Charles (2010) did not find evidence of a shared community of teachers working collaboratively within a whole school philosophy of formative assessment and teaching. The analysis of teaching indicated that teachers were still teaching according to the formulaic procedures that are a legacy of the Numeracy and Literacy Strategy. As such, teachers had yet to integrate the philosophy of formative assessment into their teaching and professional beliefs.

School level. The studies presented under 'School level' focus on two different types of policy. One study examines the sustainability of formative and summative assessment practices

within school policy (Jones & Moreland, 2005), while the other two studies examine the nature and effectiveness of implementation in response to national policy (Hume & Coll, 2009; Priestley & Sime, 2005). The review in this section compares the implementation of formative assessment in a school in New Zealand (Jones & Moreland, 2005) with that in a school in Scotland (Priestley & Sime, 2005). The third study which investigates formative assessment in the New Zealand science curriculum is presented on its own.

For formative assessment to thrive and be sustained beyond the initial years in a school, aspects like supportive leadership (Jones & Moreland, 2005; Priestley & Sime, 2005), new mindsets among teachers through change culture (Jones & Moreland, 2005; Priestley & Sime, 2005), and valuing and recognizing teacher voice (Priestley & Sime, 2005) are critical components of school policy. These features were encapsulated in the sustaining of a three-year program to scale up and sustain formative assessment practices in a New Zealand primary school (Jones & Moreland, 2005) and in the evolution and development of AifL in a Scottish primary school (Priestley & Sime, 2005). In the New Zealand school, the critical success factors for the continuity of the use of formative assessment practices are teachers' pedagogical content knowledge, evidence of improved student motivation, a sustained period of intervention, supportive and encouraging school culture, and effective partnership between teachers, researchers and school leaders (Jones & Moreland, 2005). An additional significant factor present in the implementation of AifL in a Scottish primary school was the alignment of school processes and procedures with the macro strategy applied in Scotland (Priestley & Sime, 2005). AifL principles were congruent with teachers' professional and personal values. Within the Scottish school, there was leadership support provided by the deputy head, a critical criterion as this support helped teachers overcome their lack of confidence in applying formative assessment

strategies. The data also indicated that the school recognized the value of incorporating teachers' voices, and took into consideration teachers' concerns about marking and thinking time. Finally, positive feedback from students increased teachers' motivation to deepen their pursuit of the AifL practices. Even so, while there were changes in classroom practices, there was also evidence of some "convergent" formative assessment practices. Overall, within this school, some teachers made more progress than their colleagues (Priestley & Sime, 2005).

The third study examines student learning, and teachers' assessment practices in relation to a new curriculum. This study of teachers' implementation of science in the New Zealand Curriculum (SiNZC) focused on *what*, *why* and *how* students were learning in science (Hume & Coll, 2009). SiNZC was introduced in the 1990s as part of large-scale educational and curriculum reforms. The data were based on interviews (with two teachers and 4-5 students from Year 11) and lesson observations in two schools. The findings of this qualitative case study indicated that teachers aligned science practices so closely to the standards and achievement level, that students ended up experiencing a narrow curriculum in which they viewed scientific inquiry as "fair testing, and … acquiring assessment techniques" (Hume & Coll, 2009, p. 286). This finding speaks directly to Torrance's (2007) concern that when teachers apply formative assessment practices closely – or convergent formative assessment – in relation to specified criteria, they end up narrowing the learning experience, and creating assessment *as* learning, instead of assessment *for* learning.

Possible explanations for these findings may be grouped under House's (1978, 1981) technological, cultural, and political perspectives of change. In these reforms, great detail is placed on developing expertise through the provision of exemplars and curricular resources (Hume & Coll, 2009)—what House (1978, 1981) calls the technological perspective. However,

culturally, teachers had yet to reconcile and align their practices with the spirit of the scientific inquiry element in SiNZC, and, as a result, their intended curricular, departmental plans, and the interpretation of the curriculum policy were very closely aligned to the "powerful effect" of the SiNZC's achievement standard (Hume & Coll, 2009, p. 284), rather than to the essence of scientific inquiry, which the researchers define as involving open-ended problem-solving and the use of diverse methods.

Classroom level. The effective implementation of formative assessment in the classroom transcends teachers' judicious planning of lessons. Instead, teachers need to be able to reconcile their existing beliefs, routine practices, and values with the reform before being able to embrace the spirit of the policy. This complexity may sometimes be exacerbated when a western assessment reform is introduced into a Chinese society, as evidenced in a study which examines the way three Hong Kong primary teachers implemented the *Assessment for Learning* policy into their classrooms (Forrester & Wong, 2008). Based on lesson observations, interviews, teacher reflections, and analysis of student work, the findings indicated that teachers' use of peer and self-assessment improved students' personal and social development (Forrester & Wong, 2008). However, because teachers continued with their traditional "teacher-mediated" practices, rather than exposing students to the formative intent of AfL, they unconsciously directed students towards summative learning (Forrester & Wong, 2008, p. 282). In rationalizing this behavior, Forrester and Wong (2008) suggest that these teachers' decisions to enact more teacher-centered strategies resulted from an interaction between western pragmatism, and eastern Confucianism, and Daoism—three philosophies which symbolize the complex nature of Hong Kong society. Based on the findings, it is therefore important to engage with teachers' entrenched mindsets and views before initiating reforms (Forrester & Wong, 2008).

Summary. A change in classroom practices is effective only if long term sustainability and scalability is guaranteed. The findings from the studies presented in this category describe the types and quality of change in teachers' formative assessment practices at the national, school, and classroom levels. The New Zealand (Jones & Moreland, 2005) and Scotland (Kirton, et al., 2007; Priestley & Sime, 2005) national and school-level studies demonstrate that for successful change in classroom assessment, engaging with teachers is critical since it is they that need to incorporate and understand the policy. Learning from the Scottish experience, at both the national and school levels, effecting system-wide and sustainable changes in classroom practices, requires concerted efforts to be made structurally (e.g., comprehensive support structures), culturally (e.g., teacher involvement in the conceptualization), and politically (e.g., empowerment of schools to customize the change process according to specific local conditions). One aspect missing in these studies is the use of student achievement data to illustrate the impact of formative assessment. The New Zealand study which sought to examine student learning in the SiNZC curriculum only provided qualitative descriptions of student learning. As such, it was not possible to conclude if the scientific inquiry curriculum had a positive impact on students.

Summary of conceptual and empirical work on formative assessment

Since Black and Wiliam's (1998a) seminal review, the formative assessment field has continued to evolve, with scholars engaging in inductive and deductive theory-building work, drawing from diverse fields including learning theories, critical theory and activity theory. Compared to authentic assessment, formative assessment has witnessed a tidal surge of interest, as evident in the different ways in which this concept has been theorized, and in the numerous frameworks that have been advanced over the same time period.

At the same time, like authentic assessment, the formative assessment field continues to struggle from a lack of unique identity. Recent efforts by Black and Wiliam (2009, 2012b) and Wiliam (2010) to advance a unifying model and theory of formative assessments attest to the existence of a still disparate field. Interestingly, there are different views on the existing state of the field. Brookhart (2007, p. 47) asserts in her review that it is no longer true that there is a "paucity" of research in formative assessment, while Bennett (2011, p. 21) claims that the field is still a "work-in-progress" and urges scholars to "continue the hard work needed to realize its considerable promise." One observation is that scholars recognize the value of both formative and authentic assessment. However, in the theoretical framework, neither group of scholars has integrated these two fields, except for Wiggins (1989) who spoke of the value of formative assessment as part of the authentic assessment process. Nevertheless, among the empirical work, there were two studies (Fox-Turnbull, 2006; MacPhail & Halbert, 2010) that have looked at the impact of formative assessment on authentic tasks.

The empirical work presents formative assessment as a complex field. Based on the studies under *Pedagogue*, it is evident that teachers value formative assessment practices and that they have some understanding of the philosophy and intent. However, their practices are limited, possibly because they lack the skills to use formative assessment strategies, are hindered by the conflicting policy messages, have students who are unfamiliar with the nature of the formative assessment classroom, or have unsupportive colleagues and administrators.

Intervention studies formed the bulk of the studies, especially under *Procedures and tools*, indicating researchers' deep interest in harnessing the promise of formative assessment. To this end, the review presented many studies which featured strategies, tools, and methods to improve formative assessment practice. The findings provide conclusive evidence of the impact of

formative assessment. Based on the interventions, all but two studies (e.g., Yin, et al., 2008) reported that formative assessment has positive impacts on student achievement (e.g., Black, et al., 2003b, 2004).

The studies under *Procedures and tools*, and *Student* provide suggestions for educators, researchers, and policy makers as to how to best harness the value of formative assessment. Under *Procedures and tools*, the findings indicate that there are benefits from specially defined strategies (e.g., Ruiz-Primo & Furtak, 2006b) and manipulatives (e.g., MacPhail & Halbert, 2010). However, the effectiveness of the use of these aids requires teachers to be strong in their pedagogical content knowledge (e.g., Aschbacher & Alonzo, 2006), and to recognize the informal formative assessment tools (e.g., Leat & Nichols, 2000). Finally, the research also identifies the importance of including student comments on and perceptions of formative assessment. Though young, the students in these studies have clear ideas of what strategies meet their needs (Williams, 2010), and also how they would like the strategies to be used (Colby-Kelly & Turner, 2007).

The findings of the studies presented under policy *Policy* are significant in pointing out the complexity of introducing and implementing changes in teaching and learning through formative assessment. These complexities occur at the national, school, and classroom levels. They demonstrate that to effect such change requires more than resourcing and capacity building. Important considerations to incorporate in the change process are to be cognizant of local conditions, to provide opportunity and time for teachers to understand and reconcile the purpose of the change with their own values and beliefs, and to include teachers in designing and developing the change agenda.

However, the findings from these studies need to be interpreted cautiously. The small samples used in many of the studies do not support the generalization of the findings to other contexts and groups. In addition, the participants and schools recruited for these studies were already formative assessment converts. There were also issues with the research design, since the majority of the studies do not employ comparison groups. As a result it is difficult to establish a causal effect due to the use of formative assessment.

Conclusion and discussion of literature review

The literature review has provided a broad overview of constructivist learning theories, the theoretical perspective which directs this dissertation's research design. Specifically this dissertation uses constructivist learning theories to examine and analyze the *authentic intellectual work* and *formative assessment* concepts, which together, provide the framework for analyzing the teaching of the knowledge and skills espoused in Singapore's TSLN-TLLM visions. Scholars working in *authentic intellectual work* and *formative assessment* draw on constructivist theories, yet, there are few studies that link the two. I suggest that authentic assessment provides the architecture to design assessments that elicit higher-order thinking and application skills deemed necessary in Singapore's TSLN-TLLM vision, while formative assessment provides a means for teachers to make sense of students' work, and to take steps to help them enhance their learning.

The review of the conceptual work on authentic assessment indicates a still evolving field, with persisting debates on the definition of the term, and with theorists contributing from a variety of fields in recent years. Yet, as evident in the international empirical work on authentic intellectual work, scholars have continued to adopt and adapt Newmann and Associates' (1996) framework. While additional standards and criteria have been used to sharpen the conception of

authentic assessment based on different contexts, the basic three criteria – *Construction of Knowledge*, *Disciplined Inquiry*, and *Value Beyond School* – proposed by Newmann and Associates remain the cornerstone of many frameworks. Additionally, while some scholars (e.g., Ben-Chiam, et al., 2007) do not refer to Newmann's framework, their conception of authentic assessment has similar features to the three criteria. This suggests that despite the polemical discussion among philosophers (e.g., Splitter, 2009), there is a global consensus on the essential features of authentic assessment.

In the first decade after Newman and Associates introduced the AIW concept, most of the empirical work was dominated by large-scale studies led by university researchers. Authentic assessment was not the variable of interest in these studies. Instead, these studies used authentic assessment as a means of analyzing and examining the quality of teaching and learning as a consequence of school reform efforts. However, in the second phase, beginning in 2004, there have been a number of studies (e.g., Dennis & O'Hair, 2010; Gulikers, et al., 2004) that position authentic assessment as the key variable of study. This suggests an increased interest in the field, as well as the development of different interventions to examine this concept.

There are several issues emanating from the empirical research. The large-scale university-led studies collected many pieces of student work, but did not follow through on specific students. The high rates of student mobility in countries like the USA and Australia, as well as the changes in course schedule, prompted teachers to submit work from different students and classes. In CORS-SRS, Newmann et al. (1996) reported that in their sample, just 45% of the work was collected from the same students. In addition, in these educational jurisdictions, there is no common curriculum and thus, it is difficult to establish the quality of teacher assessment and student work. In terms of the subjects assessed, the studies in the empirical work examined

mostly core subjects in the curriculum, and geography or earth science has yet to be incorporated in any of the large or small-scale studies. This subject has the great propensity for teachers to design tasks aligned to Newmann and Associates' (1996) framework. Yet, it is possible that teachers' tasks do not elicit higher-order thinking. To this end, this dissertation seeks to address this gap in the literature. First, the dissertation engaged teachers in dialogue to ascertain their views of "assessment" before examining the quality of assessments that they design. Second, Singapore has a national geography curriculum which depicts the knowledge, skills, and values that are to be delivered. This provides a common yardstick for analyzing the quality of geography teachers' assignments over a five-month period.

The development and evolution of conceptual and empirical work on formative assessment indicates a more sophisticated field, given the number of theoretical and empirical pieces that were identified for review. In comparison to authentic assessment, formative assessment appears to be a more unified field, as indicated by the number of research reviews and theoretical papers. In addition, theorizing in formative assessment has developed rapidly, with models describing practice (e.g., convergent and divergent assessment practices in Torrance & Pryor, 1998) and theorizing that draw on other perspectives (e.g., Pryor & Crossouard, 2008, take reference from socio-cultural theory and activity theory). However, it is also a disparate field because a comparison of the empirical work included in each review indicates that different studies are included. To this end, the field requires more theory-building (Bennett, 2011).

The research questions, methods, analyses, and findings from the empirical work on formative assessment indicate a vibrant field, with observational and intervention studies examining the use of formative assessment at the personal, classroom, school, and national levels. In the majority of the studies presented, student learning associated with teachers' use of

formative assessment is measured by improvement in test scores. Only a few studies examined other learning outcomes like understanding (Davies, et al., 2004), self-esteem (Miller & Lavin, 2007), self-assessment (McDonald & Boud, 2003), and the application of learning to a new context (Fox-Turnbull, 2006). Thus, one gap which this dissertation seeks to address is to examine how teachers use formative assessment to enhance student learning in higher-order skills, as manifested in Newmann and Associates' (1996) AIW criteria.

Of the 53 studies reviewed, only four studies presented under "Decisions" (Coffey, et al., 2005; Gattullo, 2000; James & Pedder, 2006; Tiknaz & Sutton, 2006) focused on the decisions teachers made based on the interpretation of student work. Given that the intent of formative assessment is for teachers to make decisions and plan the next steps for instruction (Black & Wiliam, 2009, p. 9), there appears to be a gap in the literature in terms of eliciting and understanding the follow up action that teachers would take. To this end, this dissertation addresses this gap by inviting teachers to discuss their interpretations of student work, and to identify the approaches that they would employ to improve student learning.

Finally, this dissertation integrates authentic assessment and formative assessment. To this end, the research method involves teachers discussing how they interpret student responses to the tasks they design, and in suggesting how they apply formative assessment to enhance and improve student learning. The aim is to build on the developmental nature of learning in a way that is congruent with constructivist learning theories. To present an overview of teachers' classroom assessment, this dissertation draws on the self-report responses from five cycles of TIMSS Teacher Questionnaires (1995, 1999, 2003, 2007, and 2011) to obtain a broad pattern of Singapore teachers' assessment practices and also to provide a picture of Singapore's student learning, based on TIMSS achievement data. Second, this dissertation examines the nature and

pattern of teachers' assessment practices, and uses these as indicators of the quality of assessments presented to students. And third, the dissertation examines teachers formative assessment practices used to enhance and support student learning. These three aspects differentiate this dissertation from the earlier classroom assessment study by Luke et al. (2005). This study also adds to the research by drawing on the AIW criteria and formative assessment to examine Singapore teachers' classroom assessment practices for geography.

CHAPTER 3: RESEARCH DESIGN AND METHODS

This chapter provides the context for this dissertation study with an overview of Singapore's evolving and changing educational landscape. Next, I remind readers of the research questions, and outline the methodology and research design. The chapter concludes with a discussion of the validity and reliability issues related to the study.

Study context

The educational reform in Singapore that forms the context of this dissertation was introduced in 1997. Developing a historical perspective on study context is beneficial because it provides evidence as to whether "change efforts are sustainable achievements [or] matters of only transient interest" (A. Hargreaves & Goodson, 2006, p. 35). Knowledge of the historical, social and political underpinnings that provided the impetus for the reform will enhance understanding of the theory of action and the accompanying strategies concerning the alignment of classroom assessment practices with the policy vision.

Singapore is a small and young country strategically located at the crossroads between the West and the East, and is naturally endowed with deep sheltering waters. This made the island a natural stop-over for trading ships travelling between the colonial powers of Britain and Holland in the 19th century, and China. Following the departure of the British in 1969, this newly independent nation faced many challenges, including stimulating economic growth, enticing investors to the new industries, providing employment, and educating the population. Providing education to the population was critical because the fledging government recognized that Singapore's survival depended on the talent and skill of its workforce, given that the country is not richly endowed with natural resources. To this end, education was, and continues to be, the principal means of "retooling the productive capacity of the system" (Gopinathan, 2007, p.

59). Educational policies were, and still are, designed to meet the economic and social demands of the time. Since achieving independence from the British, the education system has played a pivotal part in Singapore's nation building (Yip, Eng, & Yap, 1994). From the early days, Singapore's leaders have carefully crafted strategic educational policies to transport a country—without compulsory education and with just a small number of high schools—"from third world to first" within one generation (K. Y. Lee, 2000).

Since 1969, education in Singapore has served the twin pillars of "developing the individual and educating the citizen" (Teo, 1998, para 19). The close alignment between Singapore's nascent education system and the economy is manifested in three key phases: *survival-driven* (1959-1978), *efficiency-driven* (1979-1996), and *ability-based and aspiration-driven* (1997 to 2011)[6] (OECD, 2011). In each phase, the goal of education was to enable Singapore to compete in the world market (Yip, et al., 1994). TSLN was launched during the ability-based and aspiration-driven phase—the third episode in Singapore's education journey.

In the survival-driven phase, the government set out to expand education facilities to increase student enrolment. This was to raise the education level of Singaporeans, a large proportion of which were illiterate and unskilled, thereby enabling them to function productively in the growing industrial sector. At that time, the curriculum had a technical bias (Yip, et al., 1994). The success of the massive expansion phase is clear: between 1970 and 2004, basic literacy rates increased from 68.9 per cent to 94.2 per cent, while the proportion of university graduates in the population rose from 1.9 per cent to 12.1 per cent over the same period (Luke & Hogan, 2006, p. 175). This education objective served its purpose until other countries in Asia

[6] The Minister for Education, Mr Heng Swee Keat, launched the fourth phase, *Student-centric, values-driven education*, in 2011. See Heng, S.K. (2011). *Opening Address by Mr Heng Swee Keat, Minister for Education, at the Ministry of Education (MOE) Work Plan Seminar, on Thursday, 22 September 2011 at 10.00 am at Ngee Ann Polytechnic Convention Centre*. This fourth phase continues to work towards the TSLN vision.

started industrializing, and by the late 1970s, Singapore was losing its comparative advantage in the low-cost assembly-line, Fordist-type industrial production economy. It was evident that there was a need to move to a higher-skilled economy (OECD, 2011).

While the survival-driven phase aimed to expand basic education, the quality of education was wanting, and there was a high dropout rate. To remedy this, a review of the education system was conducted, following which a landmark report (Dr Goh Keng Swee and the Education Study Team Ministry of Education, 1978) ushered in the second phase in Singapore's education system. This phase focused on efficiency-driven processes and outcomes where students were tracked and promoted based on their academic performance in the core subjects (Gopinathan, 1999). This reform paralleled the economic aims of transforming Singapore from a labor-intensive to a capital and skill-intensive country. The purpose of education during this period was to shift Singapore from the previous "one-size-fits-all," mass education model to one that created multiple paths for students. The goal of streaming students was to reduce the high dropout rates and to the improve the quality of labor for the workforce (OECD, 2011). Although the streaming of students had unpleasant social and academic consequences (especially for students who were late developers) and therefore received unfavorable public reaction (OECD, 2011), this educational reform succeeded in dramatically reducing the dropout rate. By 1986, only 6 per cent of students had less than ten years of formal education. It was during this phase that Singapore's students' performance in TIMSS put the country on the education world map.

The efficiency-driven phase reached its sunset at the same time as the 1997 Asian financial crisis broke out. While the strategies adopted in the efficiency-driven phase were able to meet the key objectives of a capital and skill-intensive economy, the 1997 Asian financial

crisis and the emergence of a global knowledge economy necessitated a "paradigm shift in Singapore's education system towards a focus on innovation, creativity and research" (OECD, 2011, p. 162). It was becoming evident that Singapore's obedient workforce, compliant and working efficiently under orders from the top, needed new attitudes and mindsets to compete with and to face challenges ahead. For Singapore, as with other post-industrial societies, the new global economic order is dominated by service- and knowledge-oriented industries which demand more than rudimentary skills. Workers need to be able to think independently, develop strategies to solve problems, and adapt quickly and flexibly in volatile situations (Luke & Hogan, 2006). To create the 21st century employee, Singapore entered its third phase, ability-based, aspiration-driven education. This phase is encapsulated under the *Thinking Schools, Learning Nation* (TSLN) vision. In brief, *Thinking Schools* were given more autonomy to serve as "crucibles for questioning and searching, within and outside the classroom" (C. T. Goh, 1997, paragraph 22). Schools were to be the milieus that would "fire in … students a passion for learning (C. T. Goh, 1997, paragraph 21). *Learning Nation*, the second half of the vision, calls for learning to transcend schools and educational institutions. It envisages learning that takes place at every level of society.

Devolving autonomy to schools and encouraging the bubbling of ideas from the ground up marked a turning point in educational policy, given that the Singapore government is generally perceived as 'patriarchal.' Paradoxically, the approach chosen to prepare citizens for an unpredictable and rapidly changing future was to revisit the "fundamentals of education," focusing on "holistic development in the moral, cognitive, physical, social, and aesthetic" aspects (S. Y. Tan, 2000, p. 484).

On the one hand, the TSLN approach aimed at preparing Singaporeans for the future; on the other, it was the government being cognizant that the system had for too long emphasized the acquisition of factual knowledge (Teo, 1998). To this end, the tenets of the ability-based, aspiration-driven education phase were a reduction of factual content in the curriculum to create time for group and project work, the development of Information Technology (IT) skills, and the nurturing of a passion for lifelong learning. Since many parents, students and teachers perceive of assessment as driving the curriculum (K. Tan, 2008), the Ministry of Education made fundamental changes to assessment in order to align it with the TSLN vision. Increasing the weighting of higher-order thinking skills over discrete factual knowledge, the use of alternative assessments in several subjects for the GCE 'O' and 'A' level examinations, and the adoption of new modes of assessment signaled a shift from the traditional pen-and-paper examination to other modes of assessing student learning. School-based assessment and coursework were also introduced (Y. K. Tan, et al., 2008). Another fundamental shift was the introduction of group project work which was assessed in schools, as part of the General Certificate of Education 'Advanced' Level (GCE 'A' level) examination, an assessment which is used for certification and university placement. The nature of this new assessment, among others, and the aim of the new curriculum signaled a move toward inter-disciplinary thinking, collaborative skills, independent thinking, and communication skills in which students had to develop and present their ideas and arguments cogently (Teo, 2002).

A significant lever in realizing TSLN began in 2004 when the Prime Minister called for teachers to "teach less" so that students could "learn more" (this later became Teach Less, Learn More or TLLM). The implications of TLLM on curriculum, assessment and teaching are outlined under the sections on the theoretical and conceptual frameworks. Policymakers explain

TLLM as a return to the fundamentals of teaching, focusing on the *what*, *why*, and *how* of teaching (MOE [Bluesky], 2005). Drawing from the Prime Minister's comment that "grades are not the only thing in life," TLLM reminds educators that the purpose of education is to provide young Singaporeans with "a quality of education that will prepare them for life, much more than prepare them for examinations" (Shanmugaratnam, 2005b).

Over the span of Singapore's three educational phases, the emphases have changed. But some implementation approaches remain similar, in particular, that the ground (i.e., schools) should take some ownership of the change process. In the landmark Goh report (1978), the review team pointed to the value of giving more decision-making power to the grassroots level, saying that, "[a]lthough top-down initiation has been useful, the middle and ground levels should contribute more than what they have been contributing" (para 5-1). In a similar fashion, TLLM advocated "top-down support for ground-up initiatives" (Shanmugaratnam, 2005b). In view of greater diversity and complexity in schools, this implementation strategy devolves reform to schools. The approach signals that the period of "large fixes" is over (Shanmugaratnam, 2005b). To this end, while MOE continues to shape the compelling vision, in order for the system to be nimble and responsive to students' needs, changes to teaching and learning must be driven by schools. This grassroots approach to change reflects the MOE's recognition that through their daily interactions with students, teachers are in the best position to determine what is most appropriate for the learners in their classroom.

While many reform strategies begin by modifying or creating structures such as adjusting the length of a teaching period, the implementation of Singapore's TLLM embarked on a different implicit theory of action, one of changing culture. The TLLM movement created and adopted its own change language. Everyone, from the Minister and senior policy officials to

principals and teachers in schools, spoke this change language, which used images and metaphors of 'nature' to encapsulate the vision and processes of education: learners were 'nurtured' rather than 'trained,' learners would be assisted to reach different mountain "peaks of excellence" (Shanmugaratnam, 2005b). A decade and a half since TSLN was launched, the critical question is whether teachers are enacting classroom assessment practices aligned to the policy vision espoused during Singapore's third phase of education. Basically, it is necessary to ascertain as to whether or not the reform impacts the activities and interactions among teachers and students in the classroom (Elmore, 2004).

Although Singapore is a young nation, its education system has stirred international interest because of its stellar performance in consecutive cycles of international studies such TIMSS and PIRLS, and in 2009 and 2012, also in PISA. The McKinsey foundation, which publishes research on comparative education systems, has listed Singapore in the "good to great" category based on the reforms enacted since 1997 (Mourshed, Chijioke, & Barber, 2010). As the performance of Singapore students has put the country on the international education map since TIMSS 1995, stakeholders – in particular, teachers – frequently question the need for constant and continued educational change and reform. In response to new or adapted policies and initiatives introduced yearly at the MOE Work Plan seminar, teachers and the public sometimes ask if there has not been too much change, and if more time should be provided for each reform to take root. More importantly, are the teachers responding to the policy rhetoric, and (gradually) finding new ways to assess their students, or do they merely sit and wait out the change? Another question points to the extent to which the educational change policies have impacted the classroom, the student, and learning. Is there evidence that shows that teachers' classroom

assessment practices are focusing on testing students' ability to apply knowledge? These are areas which this dissertation examines.

Based on this context, the overarching research question for this dissertation is:

Under an educational policy that emphasizes the preparation of students for "the test of life" instead of a "life of tests" (MOE [Bluesky], 2005), how do Singapore geography teachers elicit and enhance student learning through the ways they use classroom assessment?

The supplementary questions are:

1. From 1995 – 2011, what have been the patterns of Singapore teachers' classroom assessments?
    a. What forms of classroom assessments do Singapore science teachers report using in Secondary 2 (Grade 8) classrooms?
    b. How have the reported forms and patterns of classroom assessment changed over time?
    c. What are the associated patterns of student learning?

2. With respect to classroom assessment, how do Singapore geography teachers understand and use different forms of assessment in their teaching to address and enhance student learning?
    a. What does "assessment" mean to Singapore geography teachers?
    b. What is the nature and quality of classroom assessment that Singapore geography teachers create for their students?
    c. What is the nature and quality of work that students produce in response to teachers' classroom assessment?
    d. What is the relationship between the nature and quality of teachers' classroom assessment and student work?
    e. After implementing their classroom assessments, how do Singapore geography teachers make formative use of assessment data?

3. What factors influence the nature and quality of classroom assessments designed by Singapore geography teachers in response to the *Thinking Schools, Learning Nation* vision?

Research methodology

This section presents an overview of mixed methods research, the methodology used for this dissertation. I will discuss the philosophical underpinnings of this methodology, and articulate a definition of mixed methods research that was used to guide the design, procedures

and analyses of this dissertation. I conclude this section with an examination of the validity and reliability issues of mixed methods research.

Philosophical underpinnings of mixed methods research

Philosophical assumptions, worldviews and knowledge provide the basis for conducting research studies (Creswell & Plano Clark, 2007). The conceptual literature on research methods typically begins by examining the changing epistemologies and it outlines how these shape the types of research methods used (e.g., Bredo, 2006; Kelly, 2006; Strike, 2006). A full understanding of the philosophical underpinnings is important because mixed methods research embraces different paradigmatic perspectives. A full history of the rise in acceptance of and interest in mixed methods is provided in Creswell and Plano Clark (2007) and Teddlie and Tashakkori (2003).

The increasing popularity of mixed methods research as the "third methodological movement" (Teddlie & Tashakkori, 2003, p. 24) may be credited to scholars like Kenneth Howe (1988, 1992) who are the vanguard in dismissing the quantitative-qualitative debate in favor of the "critical educational research model." In so doing, Howe (1988, p. 14) advances the value of pragmatism in research, saying that "truth is 'what works.'" Scholars who argue against the "incompatibility thesis" (Howe, 1988) take the view that the differences between the qualitative and quantitative traditions are embellished while undervaluing the commonalities (Lund, 2005).

Mixed methods approaches, undergirded by pragmatism, concerned themselves with the consequences of research, adopt problem-centered approaches, embrace pluralistic perspectives, and are oriented toward real-world practice (Creswell & Plano Clark, 2007). They build on the work of scholars like Howe, Rallis and Rossman, and Tashhakkori and Teddlie who propose pragmatism as the best paradigm for advancing the use of mixed methods research. Pragmatism

is appropriate because it rejects the "incompatibility thesis" (Howe, 1988, p. 10), uses the research question to drive the research, does not require a "forced choice" between postpositivism and constructivism (Teddlie & Tashakkori, 2003, p. 679), and eschews the overuse of metaphysical concepts . Pragmatism, therefore, "presents a very practical and applied research philosophy" (Teddlie & Tashakkori, 2003, p. 21). For all the above reasons, and because this dissertation has multiple research questions that drive the research methodology, pragmatism has been adopted as the overarching paradigmatic lens for this dissertation.

Definition of mixed methods research

The most comprehensive effort to advance a definition of mixed methods research comes from Johnson, Onwuegbuzie, and Turner (2007) whose conception is based on a review of the literature and interviews with leading methodologists like John Creswell, Valerie Caracelli, Jennifer Greene, Abbas Tashakkori and Charles Teddlie. They conclude that mixed methods research has had different nomenclatures over the years, including integrative research, blended research, triangulated studies and mixed research (Johnson, et al., 2007), and more recently, "multiple methodology" (M. L. Smith, 2006). This is a varied field, with definitions based on what is mixed, when or where the mixing occurs, the breadth of the mixed research, the purpose of mixing, and the orientation of the research (Johnson, et al., 2007). Against this "homogeneity and heterogeneity" of the field (Johnson, et al., 2007, p. 123), Johnson et al. (2007) put forth two definitions of mixed methods research: the first posits mixed methods as a type of research, and the second is a general definition.

> Mixed methods research is the type of research in which a researcher or team of researchers combines elements of qualitative and quantitative research approaches (e.g., use of qualitative and quantitative viewpoints, data collection, analysis, inference technique) for the broad purposes of breadth and depth of understanding and corroboration (Johnson, et al., 2007, p. 123).

> A mixed methods study would involve mixing within a single study; a mixed method program would involve mixing within a program of research and the mixing might occur across a closely related set of studies (Johnson, et al., 2007, p. 123).

Using mixed methods research involves more than combing two or more data sources. To encapsulate the true spirit of mixed methods research, the procedure must involve the "connection, integration, or linking" of the different strands of research methods, data sources, and research procedures (Creswell, 2010, p. 51). The working definition of mixed methods research used to guide the design of this dissertation is that advanced by Creswell and Plano Clark.

> Mixed methods research is a research design with philosophical assumptions as well as methods of inquiry. As a methodology, it involves philosophical assumptions that guide the direction of the collection and analysis of data and the mixture of qualitative and quantitative approaches in many phases in the research process. As a method, it focuses on collecting, analyzing, and mixing both quantitative and qualitative data in a single study or a series of studies. Its central premise is that the use of quantitative and qualitative approaches in combination provides a better understanding of research problems than either approach alone (Creswell & Plano Clark, 2007, p. 5).

This working definition indicates that mixed methods research is more than a methodological approach. As a "more adequate science," (M. L. Smith, 2006, p. 473) mixed methods enhances the way in which research is conducted. Its use raises the validity and credibility of inferences (Greene, Benjamin, & Goodyear, 2001; Teddlie & Tashakkori, 2003), provides for increased consciousness and diversity (Greene, et al., 2001; Teddlie & Tashakkori, 2003), leads to more comprehensive findings (Greene, et al., 2001), supports research questions that other methodologies cannot (Teddlie & Tashakkori, 2003), and allows more insightful understandings (Greene, et al., 2001).

Types of mixed method research

The research literature is replete with mixed methods research designs. Within this variety of designs, the classification of the designs is typically based on the sequence, nature, and purpose of the 'mixing.' These classifications contribute significantly to mixed methods research by creating a functional organizational structure, providing examples of research designs that are distinct from quantitative and qualitative research, establishing a common language for the field, providing paths and procedures for researchers employing mixed methods studies, and serving as pedagogical tools (Teddlie & Tashakkori, 2003).

One typology was developed based on empirical work. Greene, Caracelli, and Graham's (1989) typology of mixed methods designs, based on a review of 57 empirical studies, had five categories: *triangulation* (i.e., to ensure the convergence of different methods to answer the same research question), *complementarity* (i.e., to use the results of one study to enhance the other), *development* (i.e., to use methods used sequentially to inform one another), *initiation* (i.e., to identify contradictions, gaps, and examine questions), and *expansion* (i.e., to broaden the scope, breadth and depth of the study).

Another typology comprises coordinated, integrated, and iterative designs (Greene, 2001). *Coordinated* designs allow the different methods to retain their individual characteristics. The interaction among the approaches takes place towards the end of the study when overall inferences are made. *Integrated* mixed methods designs are conducted such that the synthesis occurs during the data gathering and analysis process, while allowing the different methodologies to preserve their respective identities. *Iterative* designs are the most commonly used, and are defined by their "back-and-forth" manner (Greene, 2001, p. 257), in which the different methods are sequentially used to gradually provide insights and understandings.

Creswell and Plano Clark's (2007) typology has four different approaches. *Triangulation* designs require quantitative and qualitative data to be collected simultaneously and accords equal status to both types of data to determine if the two databases provide similar or different findings. *Embedded* designs necessitate the concurrent collection of quantitative and qualitative data, but one form of data plays a supplementary role. The data sets are analyzed separately because they pertain to different research questions. *Explanatory* designs are sequential, typically with the quantitative data collected before the qualitative data. The latter are used to refine, extended or explicate the overall findings. *Exploratory* designs begin with the collection of qualitative data, followed by quantitative data and are used to explore a phenomenon before designing an instrument to test the situation.

Drawing on the pragmatic tradition, this dissertation has adopted Creswell and Plano Clark's (2007) explanatory mixed methods design, which uses qualitative data to explicate or elaborate on the initial quantitative results. The explanatory design is appropriate for this dissertation because the aim is to examine quantitative survey data for general patterns of teacher assessment practices, and subsequently, teachers were interviewed to obtain a more detailed picture of the types of assessments they use, their reasons for using these assessments, and their interpretation of student learning. Within the explanatory mixed methods design, the collection and analysis of data within this design is done sequentially, with the findings from the second phase used to build on or explain the patterns in the first phase. However, within the spirit of mixed methods research, I also attempt to create an 'interaction' (Greene, 2001) of the data and analyses.

Data sources

This dissertation used a mix of primary and secondary data sources. Primary data sources included teacher assessment, student work, and interviews with teachers. Teacher assessment was analyzed to examine the types of learning opportunities presented to students. The quality of student learning was gleaned from the type of work that they produce in response to the assessments assigned by teachers. Secondary data sources included teachers' self-report responses from the TIMSS 1995, 1999, 2001, 2003, 2007 and 2011 Teacher Questionnaire (TQ) surveys on their classroom assessment practices, TIMSS International Science reports for the content and cognitive domains for the same five cycles, and curricular and policy documents.

Together the primary and secondary data are used to provide patterns of teachers' classroom assessment at the *macro* and *micro* levels. Due to the small size of Singapore's education system, all schools offering a secondary school curriculum are included in each TIMSS cycle. As a result, the secondary data which drew on TIMSS data provide a *macro* picture of teachers' assessment practices at the national level. The data from five TIMSS cycles present a picture of assessment practices over time. Comparatively, the primary data were collected over a five-month period during one academic year. This data provided patterns of current assessment practices at the classroom or *micro* level.

Quantitative data sources. Survey research is the dominant method of the quantitative component of this dissertation. Broadly, survey research is used to obtain trends and patterns (Creswell, 2008) about the characteristics of individuals, groups, or organizations (Berends, 2006). For this dissertation, survey research was used to obtain the patterns of teachers' classroom assessment practices at five distinctive points in time. Within this dissertation, survey

research is placed under quantitative data sources. However, the methods of collecting survey data may also include qualitative methods like personal interviews (Fraenkel & Wallen, 2009).

The survey instrument used in this dissertation drew on Singapore teachers' self-report responses from the TIMSS 1995, 1999, 2001, 2003, 2007 and 2011 Teacher Questionnaires (TQ) on classroom assessment practices. TIMSS is an international achievement study conducted every four years since 1995, and reports the achievement of fourth and eighth grade students in science and mathematics. TIMSS is a project with the International Association for the Evaluation of Educational Achievement, an independent, international organization dedicated to improving education (Mullis, et al., 2009). In each cycle, over 60 countries globally participate in the study. All of these nations have a common purpose: to improve the mathematics and science education of students in their jurisdictions. The IEA also perceives the value in comparing education systems in terms of the organization, curricular, instructional practices, and resulting student achievement as a useful policy analysis tool (Mullis, et al., 2009).

In addition to collecting and publishing international student achievement data, each TIMSS cycle collects and analyzes extensive information from students, teachers, and principals about mathematics and science curricula, instruction, home context indicators, and school characteristics. According to the TIMSS 2007 International Science Report (Martin, Mullis, & Foy, 2008), the TQ captured information from teachers teaching a representative sample of Singapore students with regards to their background, preparation, and professional development. In terms of classroom instruction, the TQ also elicited information about teachers' practices as well as detailed information about the subject matter taught to students. Within the section on classroom instruction, there is a specific set of items in each cycle that elicits teachers' classroom assessment practices.

Students, not teachers, are the sampled population in each TIMSS cycle. Teachers responding to the TQ are linked to the students selected to participate in the study, that is to say, they are the teachers teaching the sampled students. Students are selected based on a sophisticated sampling procedure. Based on the stringent sampling criteria, the exclusion rate for each country is very low. The sampling procedure employs a multi-stage process, in which schools are first selected based on the Measure of Size (e.g., student enrollment in the target grade, number of classrooms in the target grade, etc.), Minimum Cluster Size (the ratio of the total number of students to the total number of classes for schools with more than one class in the target grade), several variables (e.g., type of school, degree of urbanization, sex of students), as well as school sampling probability and status (e.g., if the school had already been sampled for a study other than TIMSS) (Joncas, 2008).

In the second stage, the sampling framework selects classes within schools. This systematic, two-stage probability proportional-to-size (PPS) sampling technique matches the hierarchical nature of most education systems, in which classes of students are nested with schools (Joncas, 2008). All students in the class selected through the sample frame participate in the study. Singapore is an exception where in addition to classes, students within each class were sampled through a third sampling stage. On the school level in Singapore, given the small population size, all schools are included in each cycle of the study. This means that teachers teaching mathematics and science to Secondary 2 (Grade 8) students in any one of Singapore's schools have the opportunity to have their students sampled. This sampling procedure also means that students, not teachers constitute the representative sample. To this end, the analysis of the TQ survey responses is discussed at the student level.

In each TIMSS cycle, the contextual questionnaires eliciting information from students, teachers, and schools are slightly modified to be consistent with the current research. As a result, teachers' responses of their practices from each cycle were analyzed within that specific time frame. The specific items on teachers' classroom assessment used in each cycle of TIMSS have been extracted from the respective TQs and presented in Appendix 1.

Rationale for using TIMSS data. One of the key questions that this dissertation seeks to examine if and how the patterns of Singapore Earth Science (Geography) teachers' assessment practices over the period after the launch of TSLN have changed, and in particular, whether the patterns show more use of formative assessment, more focus on higher-order thinking and application questions, and more varied types of assessment rather than simply mimicking the formats that appear in the national examinations. While TIMSS has data for both Grade 4 and Grade 8, this dissertation used the Grade 8 data because this is a time at which students are not close to the key stage examinations, and so, teachers have more autonomy in curricular decisions. Furthermore, at this level, Earth Science or Geography is part of the core curriculum.

There are several reasons why it is appropriate to apply TIMSS data to an examination of Singapore's educational change since TSLN. First, at the time of this study, Singapore had participated in all cycles of TIMSS (i.e., 1995, 1999, 2003, 2007, and 2011). Since TSLN began in 1997, the teacher report data from the TIMSS 1995 survey may be used as a baseline to ascertain the state of assessment practices before the policy was introduced. As educational change takes time to permeate the system, TIMSS 1999, which is two years after TSLN's introduction, provided an indicator of early patterns of classroom assessment practices. The later cycles, TIMSS 2003 and 2007, were used to compare if there have been further changes in

practice. TIMSS 2007 and TIMSS 2011 also provide a means to compare changes before and after TLLM. Details are presented in Table 4.1 in Chapter 4.

Next, because Singapore is a small country, all schools are sampled in each TIMSS cycle. The students sampled in each cycle constitute a representative sample of Singapore's population of students at Grades 7-8, the target grades for this study. Although TIMSS only samples Grade 8 students, this dissertation samples teachers teaching either Grade 7 or Grade 8, depending on at which level the selected school covers the relevant earth science topics in the TIMSS curriculum framework. Analyzing students' achievement in each cycle provides a means by which to examine changes in the education system. As mentioned earlier, this dissertation draws on the achievement data in the cognitive and content domains from five TIMSS cycles, and uses these as indicators of the quality of student learning over time.

Since TIMSS 1995, some critics (e.g., Biggs, 1996b) argue that Asian countries only perform well in international tests because students in these countries are used to test-taking, and are well-practiced in answering objective test items which require rote learning and factual recall . This assertion is based on the fact that high performing jurisdictions like Singapore, Hong Kong, and Chinese Taipei all conduct high-stakes national examinations for placement and certification purposes. Therefore, because the curriculum frameworks in TIMSS are broader, the items are written by an international expert panel, and there are no practice items, student achievement in these international studies may be seen as an alternative way of representing student learning in the content areas. In particular, skills like the transfer of learning, the application of learned material and problem-solving—all prevalent in the literature on 21st century skills—are captured in the cognitive domains of knowledge, reasoning, and application

in the TIMSS Assessment Frameworks. Specifically, because students cannot specially prepare for these tests, the likelihood of score inflation is reduced.

Finally, data from comparative international assessments "extend and enrich" the examination of the national context because such data provide a larger context within which to interpret the national performance (Schleicher, 2010). The TIMSS data set is rich because it does not merely compute an overall achievement score, but also disaggregates the scores into different cognitive domains. Thus, it is possible to examine if the curricular and instructional changes resulting from TSLN and TLLM have impacted student learning.

Qualitative data sources. As the dominant method for this dissertation, survey research was used to answer the second and third supplementary research questions. The type of survey research used for the qualitative data component was in-depth personal interviews with teachers. In this dissertation, the qualitative data are primary data and are used to present the current picture of classroom assessment practices.

The aim of the interviews is to ascertain what is on people's minds (Fraenkel & Wallen, 2009) and to understand how respondents make meaning based on their lives, experiences, and thought processes (Brenner, 2006). The advantage of interviews is that they involve in-depth discussions with participants through which rich data pertaining to the participants' views on the research topic may be gleaned. This close personal interaction between the participant and the interviewer enables the process of making meaning (Brenner, 2006; Kvale, 1996).

Inductive and deductive approaches are commonly used to design an interview study (Brenner, 2006). Inductive approaches are associated with grounded theory and are concerned with theory generation. Comparatively, deductive approaches are structured, and researchers are more explicit about the theoretical frameworks used to direct their interviews (Brenner, 2006).

This dissertation applied the deductive approach to interviewing as it is informed by constructivist theories of learning, as well as by the concepts of authentic intellectual work (AIW) (Newmann & Associates, 1996) and formative assessment.

The recruitment of teachers for the study involved an invitation extended to principals of all secondary schools as well as to the Geography Teachers' Association of Singapore. The email asked principals to encourage experienced teachers who had taught during the TSLN period to participate in the study.

I interviewed eight lower secondary geography teachers (Grades 7 and 8). This level was selected because it matched the grade and earth science (geography) curriculum frameworks assessed in each TIMSS cycle. Given that Singapore schools have the liberty to sequence the delivery of the geography topics, I decided to interview Grade 7 and 8 teachers because the lower secondary syllabus spans two years.

To add to the extant corpus of knowledge, I focused on the earth science component of the TIMSS Science Study. Previous studies by the CORS group in the USA focused on mathematics and social studies (Newmann & Associates, 1996), and the CARP study emphasized language arts and mathematics (Bryk, et al., 2000). In Australia, the QSRLS team (QSRLS, 2001) analyzed English, mathematics, science and social studies (with an environment component) and the SIPA study (Ladwig, et al., 2007) concentrated on mathematics. In Singapore, the CRPP-CRP study (Koh, et al., 2005) examined English, social studies, mathematics, and science. As such, the nature of teachers' formative assessment and authentic intellectual work in geography had yet to be explored.

To obtain an in-depth understanding of Singapore teachers' classroom formative assessment practices, this dissertation used standardized open-ended interviews with the group of

eight teachers. Each teacher was interviewed three times, each session lasting about an hour. This enabled a detailed and thorough compilation of the teachers' thoughts about and perceptions of their classroom assessment practices. For each interviewee, the findings from each interview were triangulated with the other two interviews, and analyzed in relation to the assessment submitted. This enabled me to examine if there were consistent patterns in each teacher's practice and response.

For each interview, the questions posed to participants were in two categories. The first category of questions was thematic: Interview 1 focused on the introduction and conception of assessment, Interview 2 emphasized classroom assessment practices, and Interview 3 dealt with using, interpreting, and decision making regarding assessment data, as well as reflections on that process. In addition, at all three interviews, teachers were invited to discuss the aims, purpose, design, grading and analyses of the assessments which they submitted.

The standardized open-ended interview design (Fraenkel & Wallen, 2009) was chosen because each interviewee was posed the same basic questions in the same sequence. However, the questions were worded in an open-format. The advantage of this method of interviewing lies with the fact that interviewees respond to the same questions, increasing the comparability of responses (Fraenkel & Wallen, 2009). This method also ensures that the data are complete for each person because all the topics on the interview protocol are posed to each interviewee. This interview method also facilitates the organization and analysis of the data (Fraenkel & Wallen, 2009). The disadvantage of this type of interview is that the standardized nature of the questions may constrain the natural flow of conversation, as well as the relevance of the questions and responses (Cohen, Manion, & Morrison, 2007; Fraenkel & Wallen, 2009). As a result, I have

exercised flexibility during the interviews and modified the sequence of the interview questions based on the participants' responses.

The interview protocol on teachers' assessment and student learning is adapted from the Qualitative Case Study (QCS) project from Boston College. This project was embedded within a larger study, the Teachers for a New Era (TNE) initiative at Boston College (Cochran-Smith et al., 2009). The QCS project was a three-year study that followed the trajectory of 22 participants from the time they completed their teacher preparation through their second year of teaching. The study examined teachers' learning in coursework and fieldwork; their developing perceptions of teaching, student learning, and social justice; their teaching practices during and after teacher preparation; and their overall efforts to teach for social justice (Cochran-Smith, et al., 2009; McQuillan et al., 2009). Data for this study were collected from a number of different sources, including student work, structured classroom observations, and interviews with the participants, principals, and mentors. The interviews were designed based on a protocol that reflected the changing nature of the participants' experiences during student teaching, and subsequently, as full-time teachers (Cochran-Smith, et al., 2009; McQuillan, et al., 2009). For this study, I drew on and adapted from the interview protocols designed to elicit teachers' conceptions of student learning and classroom assessment. I was part of the QCS team for two years, and used the protocol to continue working with one of the study's participants. As such, I am familiar with the protocol and its objectives. Furthermore, I previously adapted and piloted the protocol for an independent study on classroom assessment based on Singapore's School-based Curriculum Innovation projects arising from the TLLM movement.

In addition to interview data, I also collected samples of teacher assessments and student work. Following the recommendations by the LAAMP team (Clare, 2000; Clare, et al., 2001), I

requested participating teachers to submit three assessments which they hadimplemented in their classes. The decision to collect and analyze three assessments was drawn from the work of the LAAMP researchers who reported that estimates from two assessments did not provide a robust and stable estimate of the quality of the teachers' assessments.[7] Rather, they advised that there be at least two assessments be collected from teachers in order to obtain a consistent estimate of classroom practice (Clare & Aschbacher, 2001; Clare, et al., 2001).

The interview questions were based on the assessments, and included topics such as the purpose and context of the assessment, the analysis conducted after grading the assessment, and the action that teachers would take after the assessment. Each teacher was requested to identify 12 students from the class. For each of the three assessments that these teachers presented to the class, I requested that they also submit responses from these 12 students. To ensure confidentiality, students were assigned a code number. The study proceeded with an examination of the student work from those 12 students at three periods in time when the teachers implemented the work. Student work was collected in the months of April, May and July/August 2012. These time points are significant because in the Singapore academic calendar, April and May fall before the mid-year summative assessment and July/August are immediately after students return following the mid-year break. Comparing student work and teacher assessment from these two time frames is a way of examining whether student work improved following feedback and formative action taken by the teacher. With respect to the teacher, the spacing of the interview sessions provides an opportunity for teachers to reflect on the interview process as well as in their responses, because it has been found that research interviewees may develop a "positive experience" or a "change" during the interview process (Kvale, 1996, pp. 30-

[7] As reviewed and discussed in Chapter 2.

31). A 'tool kit' was given to each teacher to guide him or her in the submission of assessments and student work (Appendix 2).

Research design and analysis

The empirical work on authentic intellectual work has been conducted largely within a quantitative tradition, using regression models to isolate the effect of authentic intellectual work on student learning, or factor analyses to validate a framework of indicators for authentic intellectual work. Some studies (e.g., Koh & Luke, 2009; Newmann, Lopez, et al., 1998) present pieces of student work and provide descriptive accounts of the quality of authentic intellectual work. With the exception of the LAAMP group (Clare, 2000; Clare, et al., 2001), teachers' voices and perspectives have not been the main areas of concern in these studies. Using an explanatory mixed methods research design anchored in pragmatism, this dissertation adopts the AIW criteria to rate teachers' assessments with the aim of understanding and analyzing Singapore geography teachers' conceptions of classroom assessment and how they seek to enhance student learning through the way they construct classroom assessment within a recent policy initiative.

Overview of analysis

Research question 1:

1. **From 1995 – 2011, what have been the patterns of Singapore teachers' classroom assessments?**
    a. **What forms of classroom assessments do Singapore science teachers report using in Secondary 2 (Grade 8) classrooms?**
    b. **How have the reported forms and patterns of classroom assessment changed over time?**
    c. **What are the associated patterns of student learning?**

To answer this research question, I used documentary analysis of quantitative and qualitative data sources. To examine the patterns of classroom assessment over time, I analyzed

teachers' self-report responses to the section on 'assessment practices' in the TIMSS 1995, 1999, 2003, 2007 and 2011 surveys. In each cycle of the study, there is a section of eight to ten survey items that ask teachers to report on their assessment practices. Additionally, documentary analysis of TIMSS International Science reports for the 1995, 1999, 2003, 2007, and 2011 cycles provided indicators of Singapore's student learning, in terms of the overall science achievement scores and of the cognitive domains of knowledge, reasoning, and application over time. The hypothesis is that in each successive TIMSS cycle, there should be an increased percentage of Singapore students achieving in the reasoning and application domains if the intent of TSLN-TLLM is being practiced or implemented in the classroom.

For the data analysis, I computed the *z*-score and examined the means and standard deviations of the responses to the survey items in each cycle to create a broad picture of teachers' classroom assessment practices. I also examined the data compiled in the TIMSS International Science Reports. Further details are in Chapter 4.

Research Question 2:

2. **With respect to classroom assessment, how do Singapore geography teachers understand and use different forms of assessment in their teaching to address and enhance student learning?**
    a. **What does "assessment" mean to Singapore geography teachers?**
    b. **What is the nature and quality of classroom assessment that Singapore geography teachers create for their students?**
    c. **What is the nature and quality of work that students produce in response to teachers' classroom assessment?**
    d. **What is the relationship between the nature and quality of teachers' classroom assessment and student work?**
    e. **After implementing their classroom assessments, how do Singapore geography teachers make formative use of assessment data?**

Sub-questions (a) to (e) are used to unpack Research Question 2. For Parts (a) and (e), I used an interview protocol to elicit Singapore Geography teachers' conceptions of authentic and

formative assessment. Parts of this protocol were piloted in an independent study which I conducted earlier and other parts were adapted from the interview protocols from Boston College's QCS, as well as protocols created by Clare and Aschbacher (2001) in connection to their analysis of teacher assessments in California.

With respect to Part (a), teachers' understandings of "assessment" were elicited through a constructivist activity. Inspired by the method E. Hargreaves (2005) used in the study reviewed in Chapter 2, I provided respondents with note cards, and requested them to write words or phrases that came to mind when they thought of 'assessment.' Following this, the grouping and categorization of the high frequency terms and words enabled me to obtain a pattern of teachers' views on assessment.

The aim of parts (b) to (d) was to obtain a picture of Singapore teachers' assessment. I collected three assessments from each of eight teachers over a span of twenty weeks (or two academic terms). To analyze the quality of teacher assessment tasks, as well as quality of student work, I applied the three AIW criteria (Newmann, 1996; Newmann & Associates, 1996; Newmann, et al., 1996; Newmann, et al., 1995) to rate teachers' assessment. Three raters, including myself, rated the assessments. All three raters are Earth Science teachers. There were several coordination sessions to unpack and understand the AIW criteria, as well as trial sessions to ascertain that all raters were applying the criteria consistently.

The rating of the student work was conducted using the same process. Student work was rated based on *Construction of Knowledge* and *Disciplined Inquiry* criteria, since Newmann and Associates (1996) did not develop standards for rating the *Value Beyond School* criterion. This was because the CORS team did not have the resources to interview students to find out if they perceived the tasks to be meaningful and purposeful as indicated in the *Value Beyond School*

criterion. Finally, correlation analyses of the ratings for teacher assessment and student work were conducted to explore if there was an association between these two variables.

Finally, for Part (e), the interview protocol was used to obtain teachers' comments on how they sought to improve or enhance student learning based on the completed pieces of student work. The responses were analyzed using theoretical coding (Maxwell, 2013), a deductive data analysis method. The theoretical underpinning was based on constructivist learning theories as presented in Chapter 2.

Research Question 3:

3. **What factors influence the nature and quality of classroom assessments designed by Singapore geography teachers in response to the Thinking Schools, Learning Nation vision?**

The answers to Research Question 3 were obtained from the semi-structured questions in the interview protocol. The interview questions relevant to Research Question 3 were asked at the final interview because one of the intentions of the interview process was to provide time and space for teachers to reflect on their assessment practices over the course of the 20-week participation in the study. They were asked to discuss the decisions they had made when designing the nature, format, and objectives of their classroom assessments. Their responses were analyzed using grounded theory (Charmaz, 2000; Glaser & Strauss, 1967) to obtain factors common to the eight teachers in explaining their assessment practices. Table 3.1 summarizes the details of the data collection, interpretation, and analysis.

Table 3.1

Summary of data collection, interpretation, and analysis methods.

Research question	Data source	Details	Data analysis
1a, b, c	TIMSS Teacher Questionnaire (documentary analysis)	5 cycles (1995, 1999, 2003, 2007, 2011)	Using descriptive and inferential statistics Documentary analysis: Extracting patterns of student achievement in TIMSS cognitive and content domains in order to examine patterns of student learning
2a, b, c, d, e	Interviews Student work Teacher assessment Policy documents	Three 45-min to one-hour interviews per teacher based on questions from the interview protocol (n=18 – 24 interview hours)	Theoretical coding to develop emergent themes examining teachers' conceptions of assessment and use of formative assessment Rating student work and teacher assessment using the AIW criteria (Newmann & Associates, 1996)
		8 teachers, each sharing 3 assessments (n=8 teachers)	Comparing and analyzing the quality of teacher assessment using means and standard deviation
		12 sets of student work per teacher per assessment; 6 sets each from what teacher deems work from high and middle ability students (n=288 student assessments)	Comparing and analyzing the quality of student work using means and standard deviation Using correlational analysis to examine the relationship between teacher authentic assessments and student work
3	Interviews	Same protocol as interviews in Research Question 2	Using grounded theory to develop and analyze emergent themes

Analysis involving interaction of quantitative and qualitative components

The analysis involved in a mixed methods study requires the researcher to be competent in employing a variety of quantitative and qualitative data analysis techniques, and in integrating the findings from both research domains (Onwuegbuzie & Combs, 2010). In fact, "effective

integration is a necessity for coherent and meaningful meta-inferences" to bring about increased V*erstehen* (or understanding) (Onwuegbuzie & Combs, 2010, p. 398). Since this dissertation has adopted a mixed methods approach, the inferences gleaned from the data as well as from the data analyses has involved the integration, mixing, and combination of both types of data sources. In the explanatory mixed methods research design, the quantitative and qualitative components are analyzed separately (Creswell & Plano Clark, 2007). However, the findings from the quantitative may influence the nature of the qualitative component. For example, when I prepared an interview protocol to elicit teachers' thoughts about their classroom assessment practices, some of the questions had to be modified given the quantitative findings.

There are several ways to integrate data and analyses in mixed methods research, among them Onwuegbuzie and Teddlie's sequential quantitative-qualitative analyses, Greene's phases of analysis, and Onwuegbuzie and Teddlie's seven-step process (Onwuegbuzie & Combs, 2010). Drawing on the different analytical models, Onwuegbuzie and Combs (2010) advanced the "cross-over (mixed analysis) strategies" which have several "mixing" strategies: integrated data reduction, integrated data display, data transformation, data correlation, data consolidation, data comparison, data integration, warranted assertion analysis, and data importation. In this framework "one or more analysis types associated with one tradition are used to analyze data associated with a different tradition" (Onwuegbuzie & Combs, 2010, p. 422). Based on the research questions, this dissertation drew on Onwuegbuzie and Combs' (2010) strategies to analyze the data. Due to the scope and scale of this study, it was not possible to cover all nine steps. As a result I employed the three cross-over analysis strategies shown in Table 3.2.

Table 3.2
Cross-over analysis strategies for mixed methods research.

Analysis step	Purpose and method[a]	Application to dissertation
Data transformation	Converting qualitative data into numerical codes that can be analyzed statistically (i.e., quantizing data)	Qualitative data (i.e., teacher assessment and student work) were assigned a numerical score by applying Newmann and Associates' (1996) AIW criteria. This enabled the computation of means and standard deviations to discuss and describe the quality of teacher assessment and student work. The numerical values were correlated to examine if a relationship existed between the quality of teacher assessment and student work.
Data comparison	Comparing qualitative and quantitative data/findings	The comparison of the quantitative (survey) and qualitative (interview and artifact) data ascertained if teachers' practices fell into the same pattern.
Warranted assertion analysis	Reviewing all qualitative and quantitative data to yield meta-inferences	The patterns found in the qualitative and quantitative data were analyzed and used to determine the extent to which Singapore teachers use their classroom practices to enhance student learning.

[a] Adapted from Onwuegbuzie and Combs (2010).

Data transformation. In the "cross-over" approach, data transformation is a strategy for integrating the quantitative and qualitative components of the study (Onwuegbuzie & Combs, 2010), an approach that this dissertation adopted for data analysis in order to operate within the spirit of mixed methods research. Miles and Huberman are credited with first introducing the term, 'quantizing' into qualitative methodology (Teddlie & Tashakkori, 2003). At the time, Miles and Huberman (1994) viewed data transformation as the process by which "qualitative information can be either counted directly … or converted into ranks or scales." Subsequently, Tashakkori and Teddlie (1998) extended Miles and Huberman's process to convert quantitative data, or 'qualitizing.' These terms have become part of the common taxonomy in the mixed methods lexicon and are defined as:

- Qualitized data: Collected quantitative data types are converted into narratives that can be analyzed qualitatively
- Quantitized data: Collected qualitative data types are converted into numerical codes that can be statistically analyzed (Teddlie & Tashakkori, 2003, p. 9).

As this dissertation required me to *quantize* qualitative data, I briefly explore the benefits and concerns associated with this concept. There are many ways in which quantitization of qualitative data contributes to a research study. For example,

> numbers are integral in qualitative research, as meaning depends, in part, on number … and [can be used to] establish the significance of a research project, to document what is known about a problem, and to describe a sample. [More significantly, numbers] are useful for showing the labor and complexity of qualitative work (Sandelowski, 2001).

Qualitizing processes like word counts are used to analyze free-flowing text and to identify patterns of ideas found in different text forms (Hesse-Biber, 2010; Ryan & Bernard, 2000; Sandelowski, 2001), verifying and testing conclusions and hypotheses (Sandelowski, 2001; Sandelowski, Voils, & Knafl, 2009), and transforming qualitative data into a form that enables assimilation of other quantitative data for statistical analysis (Sandelowski, et al., 2009). These counting processes are valuable because they contribute to the descriptive (getting the details right), interpretive (getting participants' experience and interpretations correct), and/or theoretical (coming up with an interpretation that is true to the facts) validity of the study (Sandelowski, 2001).

One quantitization technique is 'aggregation' as it supports the description and identification of patterns (Hesse-Biber, 2010). Common strategies used to quantitize qualitative data include frequency counts (Hesse-Biber, 2010) and rank scores (Hesse-Biber, 2010; Miles & Huberman, 1994). Frequency counts may be converted to percentages and used to provide the general meaning (Sandelowski, 2001) as well as overall summary statistics of the qualitative component of the study (Hesse-Biber, 2010). This procedure is not new because whenever

researchers place raw data into categories to look for overall common or irregular patterns, or when they attach codes and themes to data, they are "drawing from the numbered natured of phenomena in the analysis" (Sandelowski, 2001, p. 231). Data that have been structured are assigned rank scores so that they can be categorized within and across individual cases (Hesse-Biber, 2010).

In spite of the value of quantitizing data, there are drawbacks of which researchers should be cognizant. While numbers are associated with rigor and greater (scientific) precision, quantification can sometimes substitute "the simplification and security of numerical precision for the complication and ambiguity of narrative" (Sandelowski, et al., 2009, p. 219). Second, the value of quantification, like all research techniques, is contingent on the process and procedures adopted. There are four methods of quantification which may result in misleading inferences (Sandelowski, 2001). *Verbal* counting or the implication of numbers without actually giving any (e.g., use of some, many, and most) results in an inaccurate representation of the data. To remedy this, using actual numbers or providing a footnote explaining the meaning of the pronouns (e.g., that rarely means occurring in less than 20% of participants) are ways to represent an indeterminate quantity (Sandelowski, 2001). The pitfall of *overcounting* is that the overuse of numbers detracts from the focus of the qualitative nature of the narrative. A second form of overcounting is inappropriate counting just for the sake of it. In *misleading* counting, researchers inaccurately use quantitative terms to present qualitative data. Typically this is associated with small samples. As a guide, when working with samples of less than 25 cases, using the actual number is a more accurate method of presentation (Sandelowski, 2001). Finally, *acontextual* counting occurs when researchers neglect to provide other information about participants, events, or contexts and only assign a numerical value. The principle means to

overcome this is to include a narrative or a theoretical frame that corresponds to the salience of the contexts to which one is referring (Sandelowski, 2001).

Since this dissertation aims to identify the patterns in Singapore teachers' classroom assessment, quantitizing aspects of the qualitative data support and strengthene the analysis of the data, as well as ensure greater rigor and robustness in the development of themes. The quantitizing of qualitative data was applied to the analysis of the classroom artifacts collected from teacher participants. Adopting Newmann and Associates' (1996) AIW criteria, teacher assessments and student work were assigned a rating score. This numerical value was analyzed using descriptive and inferential statistics to examine the extent to which Singapore teachers presented students with challenging tasks. The pieces of student work completed in response to these tasks were similarly quantitized before analysis.

Quality of inferences

The validity of the data and results constitute an important component of good research (Creswell & Plano Clark, 2007). The purpose of validity is to check the quality of the data (Creswell & Plano Clark, 2007), and determine if the research design and findings are defensible by whomever conducted and applied the research (Creswell & Plano Clark, 2007; Onwuegbuzie & Johnson, 2006). The *mixed methods* approach to assessing quality applies criteria that address the whole mixed methods study. This is the approach preferred by most scholars (Onewuegbuzie & Teddlie, 2003; Tashakkori & Teddlie, 2008; Teddlie & Tashakkori, 2003), who advocate a two-stage assessment of the study and its conclusions, beginning by determining the validity of each individual component, and then subsequently, by analyzing the meta-conclusion for the combined mixed methods study. This overall inference or *meta-inference* refers to the "overall conclusion, explanation, or understanding developed through an integration

of the inferences obtained from the qualitative and quantitative strands of a mixed study" (Tashakkori & Teddlie, 2008, p. 101). There is value in reporting and discussing validity within the quantitative and qualitative components, in addition to the overarching meta-inference since the research is conducted using methods and procedures from each component (Creswell & Plano Clark, 2007).

Validity in quantitative research is well-developed (see, for example, discussions in Creswell & Plano Clark, 2007; Onwuegbuzie & Johnson, 2006; Shadish, Cook, & Campbell, 2002; Tashakkori & Teddlie, 2008; Thorndike & Thorndike-Christ, 2010). Its use in quantitative research differs somewhat from its understanding in qualitative research. Validity in quantitative studies is the basis for determining whether or not the results may be generalized to a larger population of interest, that is, whether the measurement is of high quality and actually measures what is claimed, thereby ensuring that the inferences are warranted. It is concerned with the trinity of internal and external validity, and reliability. Reliability refers to the consistency of the findings. According to Shadish, Cook, and Campbell (2002, p. 38), there are four major types of validity issues in quantitative research, namely, statistical conclusion validity (i.e., inferences about the correlation between treatment and outcome), internal validity (i.e., inferences about whether the observations between the treatment and outcome reflect a causal relationship between the two variables), construct validity (i.e., inferences about the higher-order constructs that represent sampling particulars), and external validity or *generalizability* (i.e., inferences about whether the relationships can be applied to different populations, contexts, and treatment and measurement variables). Of the four, there is a compromise between maximizing internal and external validity (Shavelson, 1996). Specifically, internal validity is obtained through laboratory conditions, where factors that threaten it are controlled as far as possible. As such, the

interpretations of the findings are valid, but are not generalized to situations outside the research conditions (Shavelson, 1996). On the other hand, to optimize external validity, the researcher works with conditions reflecting the real world, drawing on these variations and studying their effects. In this case, the inferences made may be affected by threats to internal validity (Shavelson, 1996). To this end, the balance between internal and external validity depends on the researcher's knowledge and selection of a research design appropriate for the topic to be studied (Shavelson, 1996).

As this dissertation used the TQ from the TIMSS study for the quantitative component of the study, some validity aspects to be addressed are the content and external validity. As an international study, TIMSS exercises great care in ensuring the validity of the data through the rigorous design process of the contextual questionnaires and achievement items (Martin, et al., 2008). The items are written by an expert panel and then field and pilot tested, and subsequently revised before the actual study is conducted. According to the Technical Reports, participating countries also provide feedback about the nature of the survey items. This ensures that the survey items have face validity within the local context of each country. Drawing on this database supports the external validity of the findings of the quantitative component. Additionally, given the sampling requirements stipulated by TIMSS, and Singapore's small country size, all schools have one randomly sampled class participating in the study. This means that the teacher teaching that class also participates by completing the TQ. To this end, the patterns obtained from the quantitative component of the study have external validity to the Singapore context because of the sophisticated sampling procedure utilized in the TIMSS research design.

In qualitative research, validity is based on whether the conclusions or narrative provided by the researcher, as well as the research participants, are accurate, trustworthy, and credible

(Creswell & Plano Clark, 2007) or, what Lincoln and Guba (1985) refer to as, credibility (the equivalent of internal validity), transferability (the equivalent of external validity), dependability (the equivalent of reliability), and confirmability (the equivalent of objectivity in quantitative research). Reliability in qualitative research differs from quantitative research, and applies primarily to the consistency of coding among multiple codes in the research team (Creswell & Plano Clark, 2007).

This dissertation used semi-structured interviews with eight teachers for the qualitative component of the study. I triangulated the findings and conducted member checks to strive towards *credibility* in the qualitative component, as Lincoln and Guba (1985) suggest. With member checks, I made available the interview transcripts to teachers so that they could verify the accuracy before I proceeded with the analysis. For ethical purposes, I also allowed them to withhold any undesirable references or comments by which they might be identified. Two types of triangulation methods were used. First, teachers' comments about their purpose and use of assessment were examined in comparison to the work they assigned. The convergence indicates if the teachers' assessment intent is aligned with their practice. Second, I interacted with the participants over a prolonged time period to better understand them and to build rapport. This latter approach is recommended by Lincoln and Guba (1985). For this reason, each teacher was interviewed three times regarding the way he or she approached authentic intellectual work and formative assessment. This enabled me to build rapport, and also served as a means to triangulate and establish consistency of their views.

Discussing validity in mixed methods is more complex. Teddlie and Tashakkori (2003, p. 36) point out that the existing terms associated with validity, regardless of whether they are used in qualitative or quantitative traditions, have been "overly used or misused." To this end, they

propose the use of the term *inference quality* as the equivalent of internal validity (for quantitative) **and** credibility (for qualitative). They conceive of inference quality as comprising two features, design quality (i.e., methodological rigor), and interpretative rigor (i.e., applicable to the accuracy or authenticity of the conclusions). Teddlie and Tashakkori (2003) also propose the use of *inference transferability* to represent the combination of external validity (for quantitative) **and** transferability (for qualitative). Their use of this term includes application and extension of the findings to new contexts, populations, time periods, and other methods of measurement. The summary of their proposed nomenclature for validity in mixed methods research is in Table 3.3.

Table 3.3

Nomenclature for validity in mixed method research (after Teddlie & Tashakkori, 2003).

Validity in quantitative research	Validity in qualitative research	Validity in mixed methods research
Internal validity	Credibility	Inference quality (design quality and interpretive rigor)
External validity	Transferability	Inference transferability (context, population, time, method)

Teddlie and Tashakkori's (2003) integrative framework for validity in mixed method research provides a way to unite and reconcile the disparate terms used to refer to the validity and credibility concerns in the quantitative and qualitative traditions (Tashakkori & Teddlie, 2008). It is also an overarching means to analyze the quality of the meta-inference about the phenomenon studied within the mixed methods study.

By applying this framework to the dissertation, I increased the *design quality* by ensuring that the selection of the design matches the nature of the research questions, such that the research questions drive the methodology selected. Additionally, the conduct of each of the

research components is aligned to the respective strengths of each of these components. As for the *interpretive rigor*, this dissertation drew on Teddlie and Tashakkori's (2008) concepts of theoretical consistency and interpretive agreement. In particular, it applied extant theories and empirical work when analyzing the data and making conclusions. Finally, in terms of *inference transferability*, the nature and design of the TIMSS data set enables the transferability of the patterns of assessment to other Singapore geography teachers, given that the teachers in the study teach a representative sample of Secondary 2 (Grade 8) students selected by a complex sampling procedure. Given that schools sampled for the study are listed in the Ministry of Education's master list of schools, there is the possibility of transferability of the patterns found to other school contexts. This is also supported by the fact that all schools have to adopt the same national curriculum. As such, teachers' practices are compared and analyzed within the same curricular content. This was an aspect absent in the analysis of teacher assessments and student work in the CORS, CARP, QSLRS, and SIPA studies because in the US and Australian states, there was no common curriculum. This made it difficult to establish a common yardstick of comparison.

Onwuegbuzie and Johnson (2006) propose that the term, *legitimation*, be used when referring to an assessment of the overall criteria for mixed methods study. They advance a typology for the legitimation of mixed methods; some elements they add are missing from Teddlie and Tashakkori's (2003) framework. There are nine indicators for the legitimation of mixed methods research in Onwuegbuzie and Johnson's (2006) framework: sample integration, inside-outside, weakness minimization, sequential, conversion, paradigmatic mixing, commensurability, multiple validities, and political. Of the nine, I discuss three which are germane to this dissertation.

Sample integration refers to the extent to which the relationship between the quantitative and qualitative sampling designs yields quality meta-inferences. In the sequential explanatory approach, it is useful for the participants in the qualitative study to be extracted from the same sample as the respondents answering the quantitative component (i.e., TIMSS study in the survey component) (Creswell & Plano Clark, 2007). However this was not possible, because the selection of teachers to respond to the TIMSS TQ was based on a random selection of schools and classrooms. To overcome this, and, knowing that given the small size of the country, all Singapore schools have the opportunity to participate in the study, I extended the invitation to participate to all geography teachers since the schools of each of them would have qualified for TIMSS. In this way, I tried to ensure a commonality in the sampling designs between the quantitative and qualitative components of the study.

The *conversion* criterion addresses the extent to which quantitizing or qualitizing yields quality meta-inferences. This dissertation quantitized the qualitative artifacts by assigning each artifact a rating score. This enabled the use of statistical analyses to identify patterns in the quality of teacher assessments and student work. In addition, drawing on the interview protocol adapted from the LAAMP and Boston College's QCS research, teachers were invited to discuss their assessments during the interviews. In this way, quantitizing the qualitative data is used to triangulate with the survey and interview data, and thereby supports the answering of the research questions.

Finally, the *weakness minimization* criterion refers to the extent to which the limitations from one approach are compensated for by the strengths from the other approach. In designing this mixed methods study for the dissertation, I used the qualitative component to build on the findings of the quantitative component. In the published CORS, CRPP-CRP, CARP and QSLRS

studies, the researchers collected, examined, and rated teacher assessments. These were then analyzed using descriptive and inferential statistics. While these studies were concerned with the quality of student learning, there was little examination of why and how teachers planned to use the assessments before and after students completed the tasks, with the exception of the LAAMP study. To this end, this dissertation used the database from a large-scale observational study to obtain patterns of the purposes and nature of teachers' assessments, and then employed a small-scale qualitative study to understand their professional decisions in greater depth.

Conclusion

This dissertation used mixed methods approaches because they were appropriate for the research questions in this study. The decision to use mixed methods research is consistent with the views of scholars in the field that the "primary importance of the questions asked" takes precedence over the methods (Creswell & Plano Clark, 2007). In fact, Viadero (2005) argues that mixed methods research is a necessary approach in the field of education because it "offers the potential for deeper understandings for some education research questions that policymakers need answered" (para 14). Because this study used mixed methods research, the adopted world view is pragmatism, employing deductive and inductive thinking (Creswell & Plano Clark, 2007) in order to examine Singapore teachers' classroom assessment practices.

CHAPTER 4: TEACHERS' CLASSROOM ASSESSMENT AT THE NATIONAL LEVEL (1995-2011)

Introduction

The movement towards greater accountability demands made on public schools in many countries during the 1960s led to the increased use of standardized national assessments and less reliance solely on teacher assessment for reporting student learning (Stiggins, 2002). This move beyond a sole reliance on teacher assessments was due to the variability in the quality of teacher-designed assessments (Crooks, 1988) and their grading practices (Cizek, 1997), the lack of teacher training and expertise in assessment (Bol, et al., 1998; Leighton, et al., 2010; Wiggins, 1989; Wolf, et al., 1991), and teachers' reports that they feel ill-prepared to conduct assessment activities (Plake & Impara, 1997).

In recent years, academic interest in teacher assessment has been on the rise, and this signifies an acknowledgement of the value of the assessments teachers use in the classroom (Leung & Rea-Dickins, 2007). One of the reasons for this shift is the recognition that teachers spend an inordinate portion of classroom instructional time assessing student learning (Stiggins, 1992; Suah & Ong, 2012). Estimates indicate that between 10% and 50% of classroom time is devoted to assessment activities (T. L. Good & Brophy, 2008; Stiggins, 1992, 2001). Teachers' classroom assessment serves many purposes, such as grading, identifying students' needs, motivating students, and monitoring the effectiveness of instruction (Ohlsen, 2007). Indeed, classroom assessment has a role to play in learning (Pellegrino & Goldman, 2008; Shepard, 2000). Classroom assessment practices such as the frequency of various types of assessment and the weight given to each serve as indicators of teaching and school pedagogy (Mullis et al., 2003). As a result, large-scale survey studies have been carried out to examine the patterns of teachers' classroom assessment (e.g., McMillan, 2001; McMillan, et al., 2002; Mertler, 1999,

2005; Suah & Ong, 2012) especially after the introduction and implementation of educational reforms. For example, Ohlsen (2007) surveyed mathematics teachers' assessment practices in response to the standards developed by the National Council of Teachers of Mathematics.

This recognition of the value and importance of teachers' classroom assessment also is evident in large-scale international studies like the Trends in International Mathematics and Science Study (TIMSS) which survey teachers' assessment practices in each participating education jurisdiction, and then compare the practices internationally. In addition to administering science and mathematics achievement assessments at the fourth and eighth grades, each TIMSS cycle collects contextual data associated with student achievement to examine factors such as home support for learning and school resources that are associated with student achievement (Mullis, et al., 2009). Such data capture information at the student, teacher, and school levels to identify and document educational and social contexts that can be analyzed with the intent of improving student learning (Mullis, et al., 2009). In each TIMSS cycle (1995, 1999, 2003, 2007, and 2011), there is a section in the Teacher Questionnaire (TQ) that elicits information about teachers' assessment practices associated with the teaching of mathematics and science.

Based on the data obtained from TIMSS 1995 and 1999, the Assessment Framework for TIMSS 2003 reported that across the participating countries teachers spend substantial amounts of time on student assessment in order to gauge student learning, plan and direct future learning, and provide feedback to students, teachers, and parents (Mullis, et al., 2003). In the first two TIMSS cycles, the survey items explored the weight teachers gave to different assessment types when assessing student work—namely, external standardized tests, teacher-made tests requiring explanations, teacher-made objective tests, homework assignments, projects or practical

exercises, students' responses in class, and observations of students (Mullis, et al., 2003). The survey also examined the purposes for which teachers used the information collected during assessments. In TIMSS 2003 and TIMSS 2007, the responses to the survey patterns indicated that teachers used a mix of formal (e.g., tests) and informal assessments (e.g., students' responses in class) to make important decisions regarding issues such as grading and accountability (Mullis, et al., 2009). Teachers also employed a wide range of assessment formats and assessed a range of content and cognitive skills (Mullis, et al., 2009).

This dissertation explores patterns in Singapore teachers' classroom assessments 15 years after the implementation of the *Thinking Schools Learning Nation* (TSLN) educational policy that aims to develop "thinking and committed citizens" in order to ensure continued vibrancy and growth in Singapore (C. T. Goh, 1997, paragraph 18). Some tenets of this policy were further defined under the *Teach Less Learn More* (TLLM) movement in 2004 that called for teachers to prepare students for the *test of life* and not a *life of tests*. Assessments that prepare students for the *test of life* are those that focus on the process rather than the product, encourage students to pose searching questions, emphasize more formative and qualitative assessing, and move away from formulaic responses (MOE [Bluesky], 2005). This vision of assessment resonates with features of constructivist assessments in which learners are able to reason critically, solve problems, and apply what they have learned to real world contexts (Shepard, 2000). Based on this background and context, the overarching research question in this dissertation is

Under an educational policy that emphasizes the preparation of students for "the test of life" instead of a "life of tests" (MOE [Bluesky], 2005), how do Singapore geography teachers elicit and enhance student learning through the ways they use classroom assessment?

1. From 1995 – 2011, what have been the patterns of Singapore teachers' classroom assessments?
    a. What forms of classroom assessments do Singapore science teachers report using in Secondary 2 (Grade 8) classrooms?

b. How have the reported forms and patterns of classroom assessment changed over time?
c. What are the associated patterns of student learning?

2. With respect to classroom assessment, how do Singapore geography teachers understand and use different forms of assessment in their teaching to address and enhance student learning?
    a. What does "assessment" mean to Singapore geography teachers?
    b. What is the nature and quality of classroom assessment that Singapore geography teachers create for their students?
    c. What is the nature and quality of work that students produce in response to teachers' classroom assessment?
    d. What is the relationship between the nature and quality of teachers' classroom assessment and student work?
    e. After implementing their classroom assessments, how do Singapore geography teachers make formative use of assessment data?

3. What factors influence the nature and quality of classroom assessments designed by Singapore geography teachers in response to the *Thinking Schools, Learning Nation* vision?

This chapter presents the analysis of the secondary data to answer Research Questions 1(a) through 1(c). Specifically, the five TIMSS cycles parallel the introduction and implementation of the TSLN education vision over the same period, and respond to the sub-questions under Research Question 1. Using the TIMSS database, I compare the self-report responses provided by Singapore teachers during five TIMSS cycles to present the *macro* pattern of assessment practices over time from 1995 to 2011. Because the student sample came from all Singapore secondary schools, their teachers' responses to the TIMSS questionnaires provide a *macro* or national picture of assessment practices. The associated patterns of student learning were obtained through documentary analysis of the TIMSS International Science Reports during these five cycles.

Singapore Secondary 2 (Grade 8) students' achievement in the five TIMSS cycles is taken as an indicator of the quality of student learning as intended by the TSLN vision. Specifically, overall science achievement and achievement in the earth science content area, as

well as achievement in the science cognitive domains (i.e., knowing, applying and reasoning) serve as the indicators for student learning in this chapter.

Second, the forms of classroom assessment practices that are presented to these students provide indicators of whether teachers are preparing students for the *test of life* (i.e., higher-order thinking skills such as the reasoning and analysis cognitive domains assessed in TIMSS) or for a *life of tests* (e.g., as evidenced by the reliance on and mimicking of national exams, and the focus on factual and recall questions). Teachers' self-report responses to the survey items in the "assessment" category of the TQ administered in each TIMSS cycle provide a macro picture of what teachers who taught a representative sample of Singapore Grade 8 students in 1995, 1999, 2003, 2007, and 2011 use in their assessment of physics, chemistry, biology and earth science. These four components are reported in relation to an Integrated Science subject in Singapore for each TIMSS cycle.

As this dissertation employs the explanatory mixed methods design (Creswell & Plano Clark, 2007), the collection and analysis of the secondary quantitative data preceded the primary qualitative data. Following the organization of this methodological strategy, the analysis and discussion of the data are likewise presented sequentially. Hence, Chapter 4 presents and discusses the secondary quantitative data that are used to provide a national picture of Singapore assessment practices over the 15 year TSLN period. This is followed in Chapters 5 and 6 by the analysis and discussion of the primary data that offers many insights into the assessment practices at the classroom (micro) level.

Survey details

Teacher questionnaire

The TQ administered in each TIMSS cycle provides the background to the classroom characteristics and instruction that are associated with student achievement. This questionnaire was completed by science and mathematics teachers who taught the sample of students participating in the study. The questionnaire looks at areas such as teachers' background (e.g., gender, academic majors, years in teaching), school conditions (e.g., satisfaction with the school, safety, school environment), instructional practices (e.g., use of computers, class size, homework), teaching resources (e.g., purposes of using computers, textbooks), and curricular plans (e.g., whether science topics are taught or not taught at the time of the study). The survey items on assessment focus on

- the weight given to different types of formal (e.g., use of teacher-made tests) and informal (e.g., observations of students) assessments – 1995 and 1999;
- the use of assessment information – 1995 and 1999;
- the frequency of giving a science test or examination – 2003, 2007 and 2011;
- the types of item format used – 2003 and 2007;
- the types of cognitive domains frequently used in science tests or examinations – 2003, 2007 and 2011; and
- the emphasis on different sources of information to monitor students' progress in science – 2007 and 2011.

The responses to these close-ended survey items are on a three- or four-point Likert scale. For example, in TIMSS 2011, the responses to the stem, "How often do you include the following types of questions in your science tests or examinations?" are in categories *never or*

almost never (coded 3), *sometimes* (coded 2)*, and always or almost always* (coded 1). All the "assessment" items in the TQ for all five TIMSS cycles are contained in Appendix 1.

As can be seen, there are no trend survey items on *assessment practices* for each cycle. However, some items are the same for 1995 and 1999, and 2003 and 2007, and then 2007 and 2011. Two items, *frequency of giving a science test or examination* and *types of cognitive domains frequently used* were administered in 2003, 2007, and 2011. As a result, it was possible to analyze and present the patterns for these survey items across two or more cycles.

Procedure

The analysis procedure began with an examination of the data to check for missing values. The missing data from the questionnaire responses comprised less than 10 percent. In compiling the international reports, TIMSS has procedures for handling missing data (Olson, Martin, & Mullis, 2008). When entering the questionnaire data using the WinDEM software, missing data are coded: (1) Not administered (i.e., respondents were not administered the question, e.g., the emphasis on standardized tests for Singapore teachers for 1995); or (2) Omitted (i.e., even though given the opportunity, the respondent did not answer the survey item). As a result, the method of dealing with the missing data was to use pairwise deletion in SPSS. In pairwise deletion, all cases in the data set that have values for two variables are included (Nouršis, 2008). This procedure is preferred over listwise deletion, in which all cases that have missing values for the variables are excluded (Nouršis, 2008). Since the missing data comprise less than 10 percent for a sample of over 5000 students, using the pairwise deletion method is appropriate in this instance. Conversely, using listwise deletion would result in the unnecessary loss of data. Being cognizant of how missing data are coded and handled is important for making subsequent

statistical decisions. Depending on the size of the sample, the default elimination strategy will have an impact on the subsequent statistical techniques used.

The five cycles of TIMSS data were analyzed using two statistical programs; namely SPSS and the International Database Analyzer (IDB Analyzer) software. The IDB analyzer is developed by the Data Processing and Research Centre (DPC) of the International Association for the Evaluation of Educational Achievement (IEA). This program is a plug-in for SPSS and enables researchers to combine and conduct analyses using SPSS data files from the IEA's assessment tools (Foy & Olson, 2009). The advantage of using the IDB analyzer is that it generates the syntax for SPSS that takes into consideration not only the complex sampling procedure used in TIMSS, but also standard errors, and plausible values for calculating estimates of the achievement scores (Foy & Olson, 2009).

When analyzing TIMSS data, the unit of analysis is the student because students comprise the sampled population. All analyses and interpretations are based on linking teachers to their students. To this end, and in order to analyze the teacher data, the first step was to use the "merge" module in the IDB analyzer to link the student and the teacher files for each cycle. The purpose of the "merge" module is to create datasets for analyses by combining different types of data files (Foy & Olson, 2009)—in this case, to combine the student and teacher files. Once the "merge" command has been executed, the IDB analyzer creates a new SPSS data file that includes both student and teacher data.

The next step was to recode the variables because some of the Likert-scale responses were combined into new categories. Then, the "analysis" module in the IDB analyzer was used to analyze the patterns of teachers' responses. The "analysis" module provides specially-created procedures for computing statistics from the TIMSS database and their standard errors (Foy &

Olson, 2009). To verify the accuracy of the analyses, the descriptive statistics (percentages and means) functions in the "analysis" module in the IDB Analyzer were compared with the statistics reported in the International Science Reports.

To determine if cycle-to-cycle changes in teachers' assessment practices were statistically significant, an independent t-test is typically used. However, because the t-distribution becomes increasingly normal in its shape for large samples (Shavelson, 1996), the *z*-score was used. The *z*-score is computed using the formula,

$$\mathrm{Z-score} = \left| \frac{\mathrm{Diff}}{\mathrm{SE(Diff)}} \right|$$

Where Diff = $\mathrm{Percentage}_{\mathrm{year+1}} - \mathrm{Percentage}_{\mathrm{year}}$

$$\mathrm{SE\ (Diff)} = \sqrt{SE^2_{year+1} + SE^2_{year}}$$

The computed *z*-score, or *z*-observed, was then compared with the *z*-critical for a two-tailed non-directional test at the 5% level of significance. If the absolute value of the *z*-score exceeds the Z-critical level of 1.96, then the difference between the two time periods is taken as statistically significant.

Findings

As the intent of the study is to use the data from five TIMSS cycles as indicators of changes in Singapore teachers' assessment practices, the data are presented chronologically. Analyzing the teachers' responses over the five TIMSS cycles is significant in determining whether there have been changes in Singapore's teachers' classroom assessment since the initial implementation of the TSLN vision. This provides responses to Research Questions 1(a) and 1(b). The areas of assessment practices surveyed in the five TIMSS cycles are: (1) emphasis on

informal and formal assessment, (2) use of assessment information, (3) frequency of giving a test, (4) use of different types of item formats, and (5) assessment of cognitive domains.

Table 4.1 presents the significant points in Singapore's third phase of education reform, the ability-based and aspiration-driven phase (1997-2011), and draws corresponding parallels to the five TIMSS cycles. The responses from the Singapore teachers in the TIMSS 1995 and 1999 questionnaires mark the years preceding the launch of the TSLN vision as well as its initial years. The responses for the third and fourth cycles in 2003 provide a picture of assessment practices about five to six years after TSLN was introduced. The survey data from the 2007 cycle are used as indicators of the state of practices a decade after the implementation of the policy. Finally, the responses to the 2011 questionnaire provide a picture of the current state of assessment practices towards the end of this phase in Singapore's educational reform. Additionally, I use the TIMSS cycles to divide the TSLN vision into two phases: **early phase** (1999-2003) and **late phase** (2004-2011). TIMSS cycles 1995, 1999 and 2003 are discussed as the early phase, and cycles 2007 and 2011 are analyzed as part of the late phase.

Table 4.1
TSLN reforms and TIMSS cycles

Year	Singapore Policy	TIMSS	Comments
Early			
1995		TIMSS 1995	Teachers' practices captured in TIMSS 1995 provide the existing patterns of classroom assessment before the launch of TSLN. Science syllabus in use was launched in 1993.
1997	Launch and implementation of TSLN		
1999		TIMSS 1999	This TIMSS assessment was conducted two years after the launch of TSLN. Teachers' practices during the TIMSS 1999 study provide an indication as to whether there has been any shift in assessment practices.
2001	Launch of new syllabuses in view of TSLN vision		For example, the 2001 Science syllabus adopts an inquiry approach (CPDD, 2000).
2003		TIMSS 2003	This TIMSS assessment represents the midpoint in the 15-year ability-based and aspiration-driven phase.
Late			
2005	Teach Less Learn More (TLLM) launched		TLLM was launched to further distill teaching, learning and assessment practices to realize TSLN. It called for more "formative and qualitative assessing," and less "summative and quantitative testing" (MOE [Bluesky], 2005).
2007		TIMSS 2007	Assessment practices reported in this TIMSS cycle provide an indication of shifts in the pattern following MOE's efforts to define the types of assessment envisioned in TSLN.
2006-2008	Launch of new syllabuses		The 2008 lower secondary science aims to nurture students as inquirers (CPDD, 2007, p. p. 2).
2011		TIMSS2011	This year was the culminating year of the ability-based and aspiration-driven phase. A new phase, Student-centric, Values-

Year	Singapore Policy	TIMSS	Comments
			driven Education (Heng, 2011), was announced in 2011. Assessment approaches reported in this TIMSS cycle indicate current practices.

Singapore's participation in TIMSS

Singapore has participated in several international benchmarking studies on student achievement since securing independence from the British colonial government in 1965. In the country's first attempt at international benchmarking studies in 1982, Singapore's students were ranked 16 amongst 26 participating countries (Y. K. Tan, et al., 2008). A decade or so later in TIMSS 1995, Singapore's students were ranked top in both the seventh and eighth for science (Beaton et al., 1996). Compared to the second international study, over forty countries participated in TIMSS 1995. Table 4.2 provides details on the number of Singapore students sampled in each TIMSS cycle, beginning in 1995, as well as the number of countries participating.

Table 4.2
TIMSS cycles and size of Singapore sample

Cycle	Sample	No. of schools	No. of participating countries
1995	4892[+]	137	42[a]
1999	4966	145	38[b]
2003	6018	164	48[c]
2007	4599	164	59[d]
2011	5927	165	63[e]

[+]Grade 8. TIMSS 1995 involved Grades 7 and 8 students.
[a]Source: http://timssandpirls.bc.edu/timss1995i/t95_countries.html
[b]Source: http://timssandpirls.bc.edu/timss1999i/participants.html
[c]Source: http://timssandpirls.bc.edu/timss2003i/countries.html
[d]Source: http://timssandpirls.bc.edu/TIMSS2007/countries.html
[e]Source: http://timssandpirls.bc.edu/timss2011/countries.html

As a small educational jurisdiction, the TIMSS sampling frame captured all Singapore schools in each cycle. Thus, each TIMSS cycle captures the achievement of representatively

sampled Grade 8 students in all Singapore schools. When referring to the grade level, I will use Secondary 2 as this is the Singapore equivalent of Grade 8.

Informal and formal assessment

In the history of education reform, there have been shifts between favoring the use of traditional large-scale (high stakes) external assessment and preferring individualized, small-scale authentic assessment (Yu & Frempong, 2012). These shifts are said to be in response to the previous domination of one type of assessment over the other (Black & Wiliam, 2005; Yu & Frempong, 2012) and an attempt to rectify this situation. At the classroom level, the international data collected from the countries participating in TIMSS show a different pattern: teachers adopt and use different sources of information to capture student learning rather than relying solely on data from high-stakes standardized tests (Martin, Mullis, Gregory, Hoyle, & Shen, 2000). This finding is based on teachers' responses to one survey item in the TIMSS 1995 and 1999 Teacher Questionnaire which elicits the emphasis students' teachers place on formal and informal assessment. The responses are in categories coded on a 4-point Likert scale, ranging from 1 (none) to 4 (a great deal).

According to behaviorist theories of learning, assessment is a scientific measure based on "precise standards" (Shepard, 2000, p. 6). In this sense, items like "teacher-made multiple-choice, true false and matching tests" which typically have a right/wrong response are less aligned with TSLN's vision that teachers should focus more on the process rather than the product of learning. Instead, assessments that require students to "describe or explain their reasoning" (TIMSS 1995 and 1999 survey item) and the use of informal assessments like observing students in class and focusing on student learning through student responses, resonate with the TLLM vision of more

"formative and qualitative assessment" and less "summative and quantitative assessment" (MOE [Bluesky], 2005).

The data for Table 4.3 were collected during TIMSS 1995 and TIMSS 1999. The TSLN vision was mooted and implemented in 1997. Therefore teachers' responses to the TIMSS 1995 and 1999 Teacher Questionnaires provide baseline indications of whether the patterns had changed after the introduction of TSLN.

Table 4.3
Emphasis on formal and informal assessments (1995-1999)[a]

Percentage of Singapore students taught by teachers who place "quite a lot" and "a great deal" of emphasis on the following when assessing student work			
		1995	1999
	Response categories	% (S.E)	% (S.E.)
Formal assessment			
Teacher-made short answer or essay tests that require students to describe or explain their reasoning	Quite a lot A great deal	58 (3.6) 18 (2.9)	53 (4.3) 17 (3.3)
Teacher-made multiple-choice, true-false and matching tests	Quite a lot A great deal	57 (3.6) 9 (2.2)	53 (4.2) 14 (3.1)
How well students do on projects[b] or practical / laboratory exercises	Quite a lot A great deal	58 (3.6) 9 (2.0)	54 (4.2) 7 (2.1)
Informal assessment			
How well students do on homework assignments	Quite a lot A great deal	42 (3.8) 5 (1.4)	36 (4.5) 4 (1.6)
Observations of students	Quite a lot A great deal	39 (3.4) 6 (1.8)	37 (4.1) 3 (1.4)
Responses of students in class	Quite a lot A great deal	34 (3.0) 9 (2.3)	33 (4.4) 3 (1.4)

[a]Tests conducted for the responses between the periods 1995 and 1999 were not statistically significant.
[b]Performance tasks like projects and practical exercises are placed under *formal* assessment. This is because in the interview responses in Chapter 6, the teachers discussed how they incorporated scores from these exercises as part of students' learning during the school year.

Over the period from 1995 to 1999, Singapore students were taught by teachers who employed a combination of formal and informal assessments. Based on the responses to the 1995

and 1999 surveys, over 60% of Secondary 2 students had teachers who placed "quite a lot" and "a great deal" of emphasis on formal assessments (including open-ended and multiple-choice tests, as well as practical project exercises), as well as on informal assessments. About 61% of students had teachers who said they used performance assessments like projects and lab experiments frequently. In the 1995 and 1999 surveys, 36 to 47% of Singapore students had teachers reporting that they place "quite a lot" and "a great deal" of emphasis on informal assessment such as oral responses, homework, and observations.

Use of assessment information

In Singapore, assessment information, especially that from examinations, is used for certification, for evaluating the education system, for placement into different academic tracks, and for selection by the institutions of higher learning (Y. K. Tan, et al., 2008). Broadly, these can be categorized in terms of the accountability and administrative functions. Assessment data also provide information that can be used formatively by teachers and students to guide instruction.

Table 4.4 presents the TIMSS 1995 and 1999 data for the purpose of using assessment information (e.g., providing marks or grades, giving feedback, reporting to parents) organized by accountability and instructional functions. In general, over ¾ of the sampled students had teachers who used assessment information for both the Accountability and Administrative, and Instructional Purposes.

Table 4.4
Use of assessment information (1995-1999)[a]

Percentage of students taught by teachers who reported using assessment information "quite a lot" and "a great deal" to			
		1995	1999
	Response categories	% (S.E)	% (S.E)
Accountability and administrative purpose			
Provide grades or marks	Quite a lot	47 (3.8)	48 (4.1)
	A great deal	26 (2.7)	31 (3.7)
Report to parents	Quite a lot	29 (3.2)	31 (4.6)
	A great deal	9 (1.7)	10 (2.5)
Assign students to different programs or tracks	Quite a lot	22 (3.1)	24 (4.2)
	A great deal	8 (1.8)	9 (2.5)
Instructional purposes			
Provide feedback	Quite a lot	64 (5.5)	69 (3.6)
	A great deal	22 (2.8)	21 (2.8)
Diagnose students' learning problems	Quite a lot	67 (3.0)	59 (5.0)
	A great deal	16 (2.6)	23 (3.5)
Plan for future lessons	Quite a lot	62 (3.0)	67 (4.2)
	A great deal	10 (2.1)	12 (3.1)

[a]Tests conducted for the responses between the periods 1995 and 1999 were not statistically significant.

About 75% of students in 1995 and 1999 had teachers who reported that they used assessment information "quite a lot" and "a great deal" for providing grades, which is an Accountability and Administrative function. This suggests that in the early phase of TSLN, teachers emphasized grades and marks as a means by which to gauge student learning. One reason for this, as will be discussed in Chapter 6, is that schools no longer relied on a final summative grade at the end of the school year as an indicator of student learning. Rather, a student's grade is the cumulative score of continual and semestral assessments.[8] As Secondary 2 is not a key-stage year, the scores of most students are not used for high-stakes purposes. However, as some schools use the Secondary 2 history and earth science scores to decide which

[8] The former includes a proportion of the scores from common and class tests conducted throughout the school year while the latter comprises a combined percentage of the midyear and end-of-year examination scores.

of the two subjects students will take at upper secondary, about one third of the students in both 1995 and 1999 had teachers who used assessment information "quite a lot" and "a great deal" to assign students to different programs or tracks. This emphasis on the accountability purpose of assessment data parallels the emphasis that teachers place on formal assessment, as discussed earlier. Finally, about 40% of students had teachers who used assessment information to report students' progress to parents.

Since over 70% of students were taught by teachers who used assessment information "quite a lot" and "a great deal" to provide feedback, to diagnose learning problems, and to plan for future lessons, this pattern of responses suggests that, Singapore students in 1995 and 1999 were also taught by teachers who used assessment information for instructional purposes.

Frequency of giving a test

In articulating the vision for Teach Less Learn More, Singapore's policy makers envisioned that teachers prepare students for the *test of life* and not a *life of tests* (MOE [Bluesky], 2005). Focusing on the latter means that students are given many practice tests and drills in order to achieve good scores in the summative tests (e.g., the end of the year school assessments or the high stakes G.C.E. O-level examinations). Conversely under the TLLM banner, the aim is to prepare students for the *test of life*, meaning that assessments focus on higher-order thinking, problem-solving and other skills needed when they leave school. The intent is for teachers to reduce the emphasis on providing "formulaic, standard answers" (MOE [Bluesky], 2005).

TLLM and TSLN do not provide explicit guidelines on the frequency of assessment. Presumably, if teachers are encouraged to move away from preparing their students for a *life of tests*, there should be a decrease in the frequency of assessment, and especially in test preparation. However, as discussed in Chapter 2, formative assessment can occur more frequently (Bloom, et

al., 1971; Thompson & Wiliam, 2008) so that teachers are able to identify areas in which their students need support in a timely manner. Thus, "formative assessing" which is encouraged under the TLLM banner (MOE [Bluesky], 2005), could suggest that teachers need to use more regular and frequent assessing in order to identify, in a timely manner, the areas in which students have learning gaps,. This would enable teachers to select appropriate strategies to address misconceptions or misunderstandings. Taking reference from the launch of TLLM in 2009, the Report of the Primary Education Review and Implementation Committee recommended that schools use more "bite-sized" assessments that focus on learning rather than just on grades (Ministry of Education, 2009). Based on this, I anticipate that the frequency of testing would likely increase, because teachers would be assessing more regularly to ascertain their students' learning, rather than waiting until the end of the unit or for a common test.

The trend responses on the Frequency of Testing from TIMSS 2003, 2007 and 2011 are shown in Table 4.5. For this item, the response categories were collapsed from five to three. The responses, "about once a week and "about every two weeks" were recoded and collapsed to become a new category, "more than once a month," while the categories, "a few times a year" and "never" were collapsed, recoded, and combined to create a new category, "less than once a month." The last category, "about once a month" was not recoded.

Table 4.5
Frequency of testing (2003-2011)[a]

Percentage of students whose teachers report giving a science test or examination			
	2003	2007	2011
Every 2 weeks or more	25 (2.1)	25 (1.7)	28 (1.9)
About once a month	61 (2.8)	52 (2.1)*	49 (2.5)
A few times a year or less	15 (2.0)	23 (1.9)*	23 (2.0)

[a]Extracted from Martin et al. (2004, p.320), Martin et al. (2008, p.335), and Martin et al. (2012, p.424).
*p<.05, significant difference 2003-2007.
( ) Standard errors.

Between the period 2003 and 2007, there was a statistically significant decrease in the percentage of students whose teachers reported giving tests and examinations "about once a month" from 61% to 52% ($z = 2.56$, $p<.05$). During the same period, there was a statistically significant increase in the percentage of students whose teachers reported giving tests "a few times a year or less" from 15% to 23% ($z = 3.1$, $p<.05$). These two patterns suggest that there is a decreased frequency of testing between the period 2003 and 2007. There was no statistically significant change in the frequency of testing for the period 2007 to 2011 as in 2011, about 50% of students had teachers who reported giving a test about once a month.

Types of item formats

The use of assessment modes such as multiple-choice or short-answer item types are frequently viewed as insufficient in gauging student learning (Wiggins, 1990). Such assessments are said to focus solely on assessing discrete pieces of facts or knowledge (Cole, 1990), and provide the impression that students' mastery or competence is binary, in terms of being seen as "correct" or "incorrect" (James, 2006, p. 54). Comparatively, "qualitative" assessments (Biggs, 1996a) are those that go beyond assessing students' ability to reproduce knowledge but rather require students to demonstrate their ability to construct new knowledge (Ertmer & Newby, 1993). To this end, open-ended constructed response questions or performance tasks are preferred because students have to explain their reasoning and ideas, or must communicate their ideas through extended writing (Newmann & Associates, 1996).

To signal the importance and value of cultivating and developing higher-order thinking skills when the TSLN vision was announced in 1997, significant and strategic changes were made to the end-of-key-stage assessments thereby placing the emphasis on higher-order thinking. For instance, the use of open-ended questions for the end-of-the-year key-stage assessments was

increased, and teachers were encouraged to use multiple modes of assessment that emphasized process skills (Ministry of Education, 1998). In a significant move away from relying on pen-and-paper assessments, *project work* was introduced in 1999 for the G.C.E. A-level examination (high school level) to enable students to examine inter-relationships and inter-connections of subject-specific knowledge (Y. K. Tan, et al., 2008). This assessment was unique as students were assessed individually as well as on their ability to work in a group. Project work also assessed skills like communication and independent learning—such skills are viewed as being necessary for life outside of school (Y. K. Tan, et al., 2008). In effect, the policy aimed to move towards the use of "multiple, complementary, and integrated" assessment (Lim & Tan, 1999, p. 398).

Upstream, the introduction of assessments like project work was strategic because it would impact the middle school level given the changes at the high school level. While the assessments for the key-stage examinations were changed, schools had the autonomy to decide on the assessment practices for the school's semesteral examinations at the non-key stage levels like Secondary 2. Given this freedom to plan and decide on the types of assessment, what patterns of assessments were reported?

Table 4.6 presents the patterns of item types presented to Singaporean students by their teachers in 2003 and 2007. The pattern of item formats used appears to be stable. For instance, for 2003 and 2007, about two-thirds of Singaporean students had teachers who relied on a mix of constructed-response and objective questions, that is, their teachers used a mix of open-ended and objective assessments. The second most commonly used assessment type is constructed-response—about one-third of Singapore students had teachers who used this. In 2003 and 2007, less than 5% of Singapore students had teachers who only employed objective questions.

Table 4.6
Item types used in science tests or examinations (2003-2007)[a]

Percentage of students whose teachers report using	2003[b]	2007[c]
Only or mostly constructed-response	30 (2.4)	29 (2.3)
About half constructed-response and half objective	68 (2.4)	68 (2.5)
Only or mostly objective	2 (0.5)	3 (1.0)

[a]Tests conducted for the responses between the periods 1995 and 1999 were not statistically significant.
[b]Extracted from Martin et al. (2004, p.321).
[c]Extracted from Martin et al. (2008, p.336).
( ) Standard errors.

Assessing cognitive domains

When TSLN was launched in Singapore in 1997, the aim was to prepare students to be citizens educated for the 21st century. In that same year, Singapore students were ranked first among all Grade 7 and 8 students in TIMSS 1995. Despite this creditable and commendable performance, the TSLN reforms pushed ahead to shift emphasis from factual recall to nurturing students who had "the ability to apply knowledge and to be creative and innovative" (Y. K. Tan, et al., 2008, p. 127). This paved the way for the focus on higher-order thinking skills. To this end, changes were made to curriculum and assessment. In assessment, there was to be a shift in emphasis from the assessment of recall and reproduction of factual knowledge towards higher-order thinking skills, such as the application of concepts (Y. K. Tan, et al., 2008).

The assessment of cognitive domains is included in the TIMSS achievement booklets. In TIMSS 2003, 2007, and 2011, the assessment items are designed around three broad categories: *knowing*, *applying* and *reasoning*. The target percentages for Grade 8 science are 35%, 35% and 30% respectively (the percentage breakdown for TIMSS 2011 is provided in Mullis, et al., 2009). Within each broad category is a further "division of behaviors" or "range of difficulty levels" as shown in Table 4.7.

Table 4.7
TIMSS 2011 Cognitive domains [a]

Knowing	Applying	Reasoning
• Recall / recognize • Define • Describe • Illustrate with examples • Demonstrate knowledge of scientific instruments	• Compare / contrast / classify • Use models • Relate • Interpret information • Find solutions • Explain	• Analyze • Integrate / synthesize • Hypothesize / predict • Design • Draw conclusions • Generalize • Evaluate • Justify

[a]The information for Table 4.7 is extracted from Mullis et al. (2009, pp.81-87) for TIMSS 2011; the details of domains are also available for the other TIMSS cycles.

There are parallels between the cognitive domains used in the TIMSS assessments and *authentic intellectual work criteria* (AIW) (Newmann & Associates, 1996). First, like authentic intellectual work, these cognitive domains are anchored within a discipline because in the most basic sense, schooling "should promote academic study" and students should be able to move beyond such knowledge to "criticism, testing, and development of new paradigms" (Newmann & Associates, 1996, p. 25). Second, the applying and reasoning domains, comprising 65% of the Grade 8 TIMSS achievement, are aligned with the *Construction of Knowledge* criterion in the AIW rubric. This AIW criterion requires that students demonstrate their ability to go beyond the recall of knowledge and content by learning to apply, assess, and evaluate. And third, the *value beyond school* criteria in AIW is echoed in the TIMSS cognitive domains because the assessment items require students to work on problem-solving and on drawing conclusions in new contexts and situations (Mullis, et al., 2009). In analyzing the assessment of cognitive domains, the hypothesis is that over the period 2003 to 2011, more teachers would be indicating that they assess higher-order skills like *applying* and *reasoning* as compared to the recall of facts.

The TQ for the 2003, 2007 and 2011 cycles focus on these cognitive domains. The responses of Singaporean students' teachers are presented in Table 4.8. In the later years of the

TSLN reforms, from 2003 to 2011, nearly 100% of Secondary 2 Singaporean students had teachers who "always or almost always" and "sometimes" assess knowing facts and concepts. In fact, the data suggests that there was a steady increase in the frequency of testing using questions based on knowing facts and concepts. There was a statistically significant increase in the percentage of students whose teachers reported that they "always or almost always" assessed the recall of learned content—from 52% in 2003 to 60% in 2007 ($z = 2.3$, $p<.05$), and then again from 2007 to 74% in 2011 ($z = 3.7$, $p<.05$). This increase over the same period is paralleled by a decrease in the percentage of students whose teachers reported that they "sometimes" designed questions based on knowing facts and concepts. The decline in the percentage of students whose teachers reported that they "sometimes" set questions based on knowing facts and concepts is statistically significant for the period 2003-2007 ($z = 2.27$, $p <0.05$), and 2007-2011 ($z = 3.4$, $p<0.05$). This increased frequency to which students were assessed facts and concepts appears to contradict the objectives of TSLN. In Chapter 6, the analysis of the interviews with the teachers provides some explanation for the emergence of this pattern of assessment practice.

Table 4.8

Frequency of using different cognitive domains (2003-2011)

Percentage of students taught by teachers who use the following types of questions on science tests				
		2003	2007	2011
	Response categories	% (S.E.)	% (S.E.)	% (S.E.)
Questions based on knowing facts and concepts	Never or almost never	1.4 (0.7)	1 (0.0)	0 (0.0)
	Sometimes	47 (2.6)	39 (2.4)*	26 (2.8)*
	Always or almost always	52 (2.6)	60 (2.4)*	74 (2.8)*
Questions based on the application of knowledge and understanding	Never or almost never	0.2 (0.2)	0 (0.0)	0 (0.0)
	Sometimes	28 (2.2)	39 (2.5)*	29 (2.5)*
	Always or almost always	72 (2.2)	61 (2.5)*	71 (2.5)*
Questions involving developing hypotheses and designing scientific investigations	Never or almost never	13 (1.8)	43 (2.5)	41 (2.7)
	Sometimes	72 (2.2)	51 (2.6)*	52 (2.6)
	Always or almost always	16 (2.0)	6 (1.1)*	7 (1.5)
Questions requiring explanations or justifications[a]	Never or almost never		4 (1.1)	3 (0.9)
	Sometimes		52 (2.5)	47 (3.1)
	Always or almost always		44 (2.4)	50 (3.0)

[a]This item was not among the survey items for 2003.

*$p<.05$, 2003-2007, and $p<.05$ for 2007-2011.

Higher-order skills in TIMSS are exemplified by the *applying* and *reasoning* domains. In the TQ, these skills are presented as *questions based on the application of knowledge and understanding*, *questioning involving developing hypotheses and designing scientific investigations* and *questions requiring explanations or justifications*. Of these three higher-order skills, the greatest change in the responses was for *questions based on the application of knowledge and understanding*. From 2003 to 2007, there was a statistically significant decrease in the percentage of students whose teachers reported that they "almost or almost always" assess the application of knowledge and understanding (z = 3.28, p<.05). But in the later phase between 2007 and 2011, there was a statistically significant increase in the percentage of students whose teachers reported that they "always or almost always" design questions that require students to apply knowledge and understanding from 61% in 2007 to 71% in 2011 (z = 2.96, p<.05). For the period 2007 to 2011, the increase in the percentage of students whose teachers "always or almost always" assess the application of knowledge and understanding is paralleled by a statistically significant decrease (z=2.81, p<.05) in the percentage of teachers who "sometimes" assess these skills. In general, there was an increase in the percentage of students whose teachers assessed the application of knowledge and understanding.

Comparatively, for the period 2003 to 2007, there was a statistically significant decrease in the percent of students whose teachers reported that they "always or almost always" assess *developing hypotheses and designing scientific investigations* from 16% to 6% (z = 4.23, p<.05). For the same period, there was a statistically significant decrease in the percent of students whose teachers "sometimes" assessed *developing hypotheses and designing scientific investigations* from 71% to 51% (z = 5.97, p<.05).

Finally, for the period 2007 and 2011, there was no change in the percentage of students whose teachers "sometimes" or "always or almost always" assess explanations or justifications.

Monitoring student progress

Table 4.9 presents the emphasis that the teachers of Secondary 2 students placed on sources of information used to monitor students' progress in science learning. In TIMSS 2007 and TIMSS 2011, there were three items in the survey relating to this issue. However, there were only two items that appeared in both cycles—namely *classroom tests* and *national or regional tests* as sources of information. The third item, *professional judgment*, that appeared in TIMSS 2007 was not included in TIMSS 2011; instead there was a new item that served as a source of information—*evaluation of students' ongoing work*. To this end, the data analysis only focuses on the two common items, *classroom tests* and *national or regional tests*.

Table 4.9
Sources of information to monitor student progress (2007-2011)

Percentage of students whose teachers emphasized various sources to monitor student progress			
		2007	2011
	Response categories	% (S.E.)	% (S.E.)
Classroom tests	Little and no emphasis	2 (0.7)	0.2 (0.2)
	Some emphasis	20 (2.0)	13 (2.0)*
	Major emphasis	78 (2.2)	87 (2.0)*
National or regional tests	Little and no emphasis	43 (2.6)	34 (2.7)*
	Some emphasis	24 (2.2)	22 (0.7)
	Major emphasis	33 (2.5)	44 (3.2)*

*$p<.05$

Aligned with the academic literature that speaks of the importance and value of teachers' classroom assessment (T. L. Good & Brophy, 2008; Stiggins, 1992), Secondary 2 students had teachers who valued the information obtained from their classroom tests, and used this to monitor student progress. From the responses in the 2007 and 2011 TQ, over 90% of students had teachers who placed "some" and "major" emphasis on classroom tests to monitor student

progress. For the same period, there was a statistically significant increase in the percentage of students whose teachers reported placing "major emphasis" on classroom assessments to monitor student progress from 78% to 87% ($z = 2.68$, $p<.05$), accompanied by a statistically significant decrease ($z = 2.57$, $p<.05$) in the percentage of students whose teachers reported that they placed "some" emphasis on the use of classroom tests to monitor student progress. The increased emphasis teachers placed on classroom tests mirrors the academic discourse which points to the value of teacher assessment as a gauge of student learning, given the amount of time teachers devote to formal and informal assessment during classroom activities (Stiggins, 1992; Suah & Ong, 2012).

Concurrently, as teachers continued to operate within the environment of high-stakes examinations, there was an increase in the percentage of students whose teachers indicated that they placed "major emphasis" on the use of national or regional tests. This statistically significant increase was from 33% in 2007 to 44% in 2011 ($z = 3$, $p<0.05$). Interestingly, this increase stemmed from a concomitant decrease in the percentage of students whose teachers reported that they placed "little or no emphasis" on the use of information from national or regional tests between 2007 and 2011. This is a statistically significant decrease from 43% to 33% ($z = 2.8$, $p<.05$). This pattern suggests that high-stakes examinations preoccupy teachers today even more than previously.

Student learning

According to Chris Watkins (2011), there are three broad categories that encapsulate what learning is: (1) learning is being taught, (2) learning is individual sense-marking, and (3) learning is building knowledge as part of doing things with others. The definition, "learning is being taught" is applied to Research Question 1(c): *What are the associated patterns of student*

learning? accompanying the teachers' assessment practices This definition is aligned with that of the TIMSS Curriculum model, which has three aspects—the intended, implemented, and attained curriculum (Mullis, et al., 2009). In a nutshell, the intended curriculum represents what society desires that students learn in mathematics and science; the implemented curriculum refers to what is taught in the classroom, and how it is taught; and the attained curriculum represents what students have learned as well as what they think about the subjects (Mullis, et al., 2009).

Table 4.10 presents Singapore students' overall achievement in science, based on the country's participation in five cycles of TIMSS. The achievement scores are taken as indicators of learning. In using the TIMSS achievement data as such an indicator, this set of data provides an alternative way to look at student learning as compared to the data from school and national examinations. This is because international benchmarking studies like TIMSS do not have test preparation material available to the public. Moreover, TIMSS is developed based on the collaboration of an international committee. Comparatively, because Singapore has a centralized curriculum and national examinations, teachers are able to focus explicitly on what will come up in these assessments. This is especially so because there are specimen papers available and teachers know how to prepare their students. Since students' good performance in the national examinations may be attributed to this thorough preparation, I suggest that achievement in international studies could be used as an alternative indicator of student learning. This proposition is based on the fact that students have not been drilled to respond to assessment items in international tests, and that the curriculum framework from an international study like TIMSS may not be completely aligned to the country's curriculum. As a result, students have to apply what they know and have been taught.

Table 4.10
Singapore students' science achievement (TIMSS1995-2011)

TIMSS Cycle	Science achievement score[a] (S.E.)
1995	580 (5.5)
1999	568 (8.0)
2003	578 (4.3)
2007	567 (4.4)
2011	590 (4.3)

[a]Extracted from Martin et al. (2012, p.40). This is the combined Chemistry, Biology, Earth Science, and Physics score.

Based on the TIMSS 1995, 1999, 2003, 2007 and 2011 International Science Reports, Singapore students performed commendably (Beaton, et al., 1996; Martin, et al., 2008; Martin, Mullis, Foy, & Stanco, 2012; Martin, Mullis, Gonzalez, & Chrowstowski, 2004; Martin et al., 2000). They were ranked first in science for the 1995, 2003, 2007 and 2011 cycles. For 1999, they were ranked second after Chinese Taipei. Students' achievement in TIMSS 2011 was stellar in that the score increased from 567 to 590 (Martin, et al., 2012). Of the 63 participating countries in TIMSS 2011, Singapore is one of seven countries participating that had an average science achievement higher than that in the previous cycle (Martin, et al., 2012). In addition between TIMSS 2007 and 2011, other than Quebec, Singapore was the only jurisdiction that had an increase in the overall science score due to an improvement in all four science content areas (Martin, et al., 2012). For Earth Science, which is the focus of the analysis in Chapters 5 and 6, the improvement in Singapore students' scores between 2007 and 2011 was statistically significant (Martin, et al., 2012).

To obtain another perspective on student learning, the overall TIMSS scores can be analyzed in terms of students' performance in the cognitive domains of *knowing*, *applying*, and *reasoning*. As mentioned earlier, higher-order skills in TIMSS are represented by the *reasoning* and *applying* domains. As a result, students' performance in these domains provides an indicator

of their ability to handle assessment questions focusing on higher-order thinking. Table 4.11 provides the data on performance in the cognitive domains as reported in the TIMSS 2011 International Science Report (Martin, et al., 2012). For TIMSS 2007, Singapore students scored second highest in the *knowing* domain after Chinese Taipei (Martin, et al., 2008). In the same cycle, Singapore students received the highest scores for the *applying* and *reasoning* domains followed by their peers in Chinese Taipei (Martin, et al., 2008). In TIMSS 2011, the overall increase in Singapore's science achievement is reflected in higher scores in all three science cognitive domains (Martin, et al., 2012). Interestingly, Singapore students performed relatively better in the *applying* and *reasoning* domains than in the *knowing* domain. Secondary analyses conducted by MOE attributed this to the TSLN shift towards more inquiry-based teaching and learning (Ministry of Education, 2012).

Table 4.11
Trends in achievement for science cognitive domains (2007 and 2011)[a]

	Knowing	Applying	Reasoning
2007	561 (4.9)	570 (4.5)	568 (4.5)
2011	588 (4.9)*	589 (4.4)*	592 (4.5)*

[a]Extracted from Martin et al. (2012, p.152 and p.165).
*2011 average significantly higher.
( ) Standard errors.

With respect to the overarching research question as to whether Singapore teachers are preparing their students for the *test of life* or a *life of tests*, the performance in the cognitive domains indicates that Singapore students can demonstrate higher-order skills and competencies as envisioned in the TSLN vision. For instance, achievement in the *applying* and *reasoning* domains indicates that students are able to apply knowledge and conceptual understanding in a problem situation as well as their being able to go beyond the ability to solve routine problems to being able to tackle unfamiliar situations, complex contexts and multi-step problems (Martin, et

al., 2008). Problem-solving in TIMSS 2011 assesses students' ability to find solutions, and in particular, for them to "identify or use a science relationship, equation, or formula to find a qualitative or quantitative solution involving the direct application / demonstration of a concept" (Mullis, et al., 2009, p. 84). Similarly, under "draw conclusions" in *reasoning*, the assessment framework in TIMSS 2011 examines students' ability to "detect patterns in data, describe or summarize data trends, and interpolate or extrapolate from data or given information; make valid inferences on the basis of evidence and/or understanding of science concepts; draw appropriate conclusions that address questions or hypotheses, and demonstrate understanding of cause and effect" (Mullis, et al., 2009, p. 86).

Another source of evidence that lends support to the quality of learning of Singapore students is the Test-Curriculum Matching Analysis (TCMA) conducted and compiled during each TIMSS cycle. The TCMA is undertaken to ensure that the international comparisons of student achievement are "fair and equitable" (Martin, et al., 2008, p. 465). Although the international representatives have endorsed the test questions, it is inevitable that the match between the test items used in each TIMSS cycle and the curriculum taught in the jurisdictions is uneven across all countries, and as a result, in each cycle, there are topics that are unfamiliar to students in some countries. The TCMA investigates the extent to which the assessment items developed for each TIMSS cycle are relevant to each country's science and mathematics curriculum (Martin, et al., 2008). This is a detailed exercise in which each country's subject specialist indicates whether an item used in the study is in their country's intended curriculum.

Table 4.12 presents the average percentage correct on the science items identified as being relevant to Singapore for each TIMSS cycle. What the data indicate is that for each TIMSS cycle, about three-quarters of the international items are similar to the Singapore

Secondary 2 science curriculum, with the exception of 2003 when there was an 85% convergence. Comparatively, countries like the USA, Australia and Romania had over 90% convergence between the TIMSS items and the country's curriculum. Based on this, it is reasonable to infer that the strong performance in each TIMSS cycle shows that Singapore students are able to apply and demonstrate their learning, even in test items for which they have not been prepared. Thus, the data from the TCMA provides further evidence of student learning for this sample of students.

Table 4.12

Grade 8 Test Curriculum Matching Analysis for Singapore (1995-2011)

TIMSS cycle	Total score points	Score points identified as relevant to Singapore	Percentage of score points relevant to Singapore	Percentage correct on all items
1995[a]	146	109	75	70
1999[b]	153	112	73	66
2003[c]	206	176	85	62
2007[d]	231	171	74	61
2011[e]	233	160	68	65

[a]Extracted from Beaton et al. (1996, p.B-3).
[b]Extracted from Martin et al. (2000, p.383).
[c]Extracted from Martin et al. (2004, p.414).
[d]Extracted from Martin et al. (2008, p.470).
[e]Extracted from Martin et al. (2012, p.484).

Discussion

This chapter has examined whether the TSLN educational policy in Singapore over a period of 15 years has resulted in changes in teachers' classroom assessment. Over this period, the Ministry of Education further defined and refined, articulated and re-articulated the education vision and the desired outcomes for learners. The vision and goals were explicitly defined in the Teach Less Learn More movement in 2005, at the midpoint of the policy period. The TLLM movement exhorts teachers to focus on the following aspects for assessment: (1) *more* process

and less product of learning, (2) *more* formative and qualitative assessing and less summative and quantitative testing, (3) *more* searching questions and less textbook answers, (4) *more* understanding and less information dispensing, and (5) less set formulaic, standard answers (MOE [Bluesky], 2005). To realize the goals of TSLN, teachers are encouraged to do less "telling" and more "facilitating, guiding, and modeling." To this end, the goals of TSLN resonate strongly with the tenets of constructivist assessment (Koh & Luke, 2009). Based on the responses from the teachers teaching the sample of students in an international study, what can be gleaned from the empirical evidence of assessment patterns of Singapore science teachers during the period preceding and that following the announcement and implementation of TSLN and TLLM?

In the early phase of TSLN, as reported in the TIMSS 1995 and 1999 questionnaire surveys, Singapore students are taught by teachers using a variety of assessment practices, including formal and informal assessments and higher-order cognitive domains. Singapore students also had teachers who used formal and informal assessments, employed assessment information for accountability and instructional purposes, and emphasized the use of both constructed-response and objective item types.

In the late phase of TSLN from 2003 to 2007, there were changes in the patterns in the frequency of administering a test or an examination toward somewhat less frequent testing. There was a statistically significant decrease in the percentage of students taught by teachers who gave a test or an examination monthly accompanied by an increase in the percentage of students whose teachers tested only a few times a year or less.

With regards to the assessment of different cognitive domains, between 2003 and 2011, almost all Singapore students had teachers who designed test questions that focused on knowing

facts and concepts, and the percentage of students having teachers that almost always did so increased with each assessment cycle. More students had teachers who "sometimes" and "always or almost always" assessed the application of knowledge and understanding, and who required explanations and justifications. From 2003 to 2007, there was a decrease in the percentage of students whose teachers assessed the ability to develop hypotheses and design investigations.

The patterns of student learning associated with these assessment practices were presented in Table 4.10. In the five TIMSS cycles, the performance of Singapore's Secondary 2 students in science has been creditable and commendable. Despite three cycles of curriculum change, and despite the fact that content had been reduced to limit over-teaching and to create time for self-directed learning (C. T. Goh, 1997), Singapore's Secondary 2 students have been ranked either first or second in Grade 8 science. This suggests that Singapore students are able to respond to questions for which they have not been prepared since the TIMSS achievement items are developed by an international panel.

Based on these broad patterns of assessment practices, what interpretations can be made from Singapore teachers' assessment patterns over the TSLN period? In the next section, I interpret and analyze these assessment practices using three main patterns—*variety*, *stability* and *change*—that have emerged from the TIMSS 1995, 1999, 2003, 2007, and 2011 survey responses.

Variety

First, *variety* was exhibited in the way teachers reported using a wide range of assessment practices. This pattern supports the findings reported in the extant research (Milnes & Cheng, 2008; Ohlsen, 2007; Suah & Ong, 2012). In terms of formal and informal assessments, or

constructed-response or objective questions, the responses from the Singapore science teachers indicated that they did not exclusively adopt a particular assessment approach. Ohlsen (2007) terms this use "hybridization" when describing the assessment practices of math teachers who are members of the National Council of Teachers of Mathematics (NCTM). In 1995, the NCTM had published assessment standards for mathematics teachers that encouraged the use of assessments closely integrated with instruction. Like TSLN, the NCTM standards called for teachers to use a variety of assessment formats to assess student learning, including the use of performance tasks, open-ended questions, and portfolios (Ohlsen, 2007).

However, unlike the Singapore teachers who frequently used extended writing (e.g., essay questions), the teachers in Ohlsen's study used this type of assessment infrequently. By comparison, the American teachers tended to assess their students using multiple-choice and short answer questions rather than extended writing. Comparatively, Singapore teachers' more frequent use of constructed-response tasks in their assessments was similar to that in Australia. For instance, researchers from the Queensland School Reform Longitudinal Study (QSRLS) reported that teachers required students to communicate ideas, concepts, arguments and explanations coherently (Hayes, et al., 2006). The study did not mention the use of multiple-choice questions in the assessments in Queensland classrooms.

While Singapore students had teachers who used a mix of formal and informal assessments, there is a greater emphasis on formal assessment. This pattern of responses emerging from the teachers of Singapore students in the early phase of TSLN differs slightly from a study by Suah and Ong (2012) who surveyed the assessment practices of 406 teachers attending in-service courses. The study was conducted in the context of a new national assessment system for public schools in Malaysia, a country situated to the north of Singapore.

The practices of the in-service teachers in Suah and Ong's study used informal assessments like observations and oral questioning. In terms of using formal assessments, Suah and Ong (2012) found that the Malaysian teachers focused on objective questions like multiple-choice questions and short answer questions. This pattern is similar to the responses provided by the Singapore teachers responding to the TIMSS survey. In fact, in Singapore, for the 1995 and 1999 TQ, more than 50% of students had teachers who placed "quite a lot" and "a great deal" of emphasis on teacher-made multiple-choice true-false and matching tests in the early phase. How can the balance in the pattern of use between formal and informal assessment be explained?

In the early years following an educational reform, the use of a hybrid of approaches enables teachers to select what works best in the context of their classrooms and schools, and is "less threatening" in the wake of reforms (Ohlsen, 2007, pp. 8-9). This is because using a variety of approaches does not require teachers to make dramatic or extensive changes to their repertoire. Adopting this varied approach allows teachers to employ aspects that enable their work to progress smoothly (Tyack & Cuban, 1995), and is also an indication of incremental change (Cuban, 1993; Ohlsen, 2007).

Another possible reason why Singapore teachers reported the use of a varied assortment of assessment practices is policy-related. Because Singapore's education system is centralized, curriculum and assessment had been reviewed and revised to be aligned with the TSLN vision. For example, the 2001 secondary science syllabus introduced during the early TSLN period recommended the use of theory tests, assignments, practical tests and mini-investigations (CPDD, 2000). Subsequently, in order to be better aligned with the TLLM philosophy of more formative and qualitative assessing, the 2008 secondary science syllabus provided a longer and varied list of assessment ideas for teachers—to use science practicals, projects, teacher observations,

checklists, reflections/journals, models, posters, games and quizzes, debates, dramas, show and tell, and learning trails as part of the inquiry-based classroom (CPDD, 2007). Furthermore, TLLM does not advocate an all-or-nothing approach. Instead, it calls for "more" of some and "less" of others, and the teachers' adopting a combination of assessment practices should mirror this policy. In this sense, the finding that Singaporean students had teachers who use of a range of assessment types seems to be policy-driven: they were aligned with centrally-driven changes to teaching and assessment.

Persistence

The second pattern of assessment practices emerging from the macro data is *persistence* in enacting practices that TLLM is calling for teachers to use *less* of. More specifically, over consecutive TIMSS surveys, there was no change in teachers' emphasis on formal and informal assessments. There was also the continued emphasis on the testing of facts and concepts.

One possible reason for the continued heavy emphasis on formal classroom tests (about two-thirds of students had teachers who placed emphasis on formal assessment)—a finding consistent with Ohlsen's (2007) study—and the use of questions that assess knowledge of facts and concepts is that the external examinations continue to have high-stakes for students and teachers. Consequently, teachers continue to design and administer assessments that mimic the format of the public examinations in order to ensure that their students are adequately prepared for these examinations (Suah & Ong, 2012). In examining the effect of TLLM on assessment, Singapore academic, Kelvin Tan (2008) points to a contradictory tension in the Prime Minister's National Day Rally speech which, in one breath, reminded students that "grades are important—don't forget to pass your exams – but grades are not the only things in life" (H. L. Lee, 2004). To K. Tan (2008), this "belies an implicit recognition of the adverse effects of high stakes

summative assessment" (p.250), which in Singapore, has come to dominate, control, and "shape students' experience of learning and schooling, in drastic ways" (K. Tan, 2008, p. 250).[9] Although the Secondary 2 teachers who responded to the surveys are not responsible for preparing their students for high-stakes examinations, the pressures associated with these examinations create a backwash effect that is likely to permeate the different education levels in schools. In such instances, it is possible that teachers teaching in non-tested grades must adjust their instruction and assessment in view of and in preparation for future testing (Hamilton & Berends, 2006). Another possible reason is the "implementation dip" (Fullan, 2007, p. 123) which typically occurs because the enactors of policy need time to reconcile with and incorporate the change. For instance, as the data in Chapter 6 indicate, schools compile an aggregate score to report student learning over the course of the school year. As a result, teachers understand this practice as a signal of the continued importance of formal assessment. Additionally, while TSLN and TLLM envision new ways to enact assessment, at the end of the day, teachers still have to prepare their students for the high-stakes assessments at the end of secondary school. To this end, the reliance on formal assessments is an indication of the intense preparation that teachers and students undertake before the national assessments. Finally, as I suggest in Chapter 7, the absence of a compelling theory to drive the changes in assessment is another reason why teachers continue to hold on to existing practices. Although TSLN was introduced in 1997, it was only in 2005 that the implications for assessment were highlighted under TLLM. As a result,

[9] At the time of this dissertation, there is a lively, passionate and contentious debate as to whether the Ministry of Education should discard the Primary School Leaving Examination (PSLE). This high-stakes assessment is administered at the end of primary school and the results are used for placement into different secondary school tracks, as well as for application to different schools. Parents and the public blame this assessment for causing stress to primary school children as well as for sustaining the highly profitable parallel education system—the private tuition industry. Concurrently, there is also strong support for retaining this high-stakes examination (http://www.channelnewsasia.com/stories/singaporelocalnews/view/1246327/1/.html)

without any vision to guide them, it is likely that in the early phase, teachers continued with their extant practices.

This pattern of *persistence* points to a tension for teachers in terms of their assessment practices. On the one hand, formal assessments such as national and school examinations compel teachers to adopt practices that lean towards accountability and the acquisition of marks. On the other hand, teachers' preference to create their own assessments (as evidenced by the assessments the teachers participating in this study submitted and presented in Chapters 5 and 6) speaks to a desire to use assessments that are more aligned to their learning goals for students. At the same time, the interview data (presented in Chapter 6) indicate that even at Secondary 2—a non-key stage level—school examinations carry high-stakes for some teachers, and thus explain their emphasis on formal assessments. This is particularly so because in some secondary schools, the end-of-the-year score comprises the marks from a number of assessments that students attempt throughout the school year. And for teachers in some schools, students' scores are an indicator of their effectiveness, a fact that compels them to adopt practices that are test-driven, that is to say, they prepare students for *a life of tests*. More details regarding these practices are provided in Chapter 6.

Change

The third pattern of assessment practices gleaned from the TIMSS survey responses is *change* towards the use of assessment practices that TLLM is encouraging teachers to employ *more* of. In particular, changes are seen in the responses for the frequency of testing, the assessment of different cognitive domains, and the way in which assessment information is used. First, there was some decrease in test frequency as fewer students had teachers who gave a test

about once a month, and as more students had teachers who tested them a few times a year or less for the 2003 and 2007 surveys.

The second *change* trend is the increased percentage of students whose teachers continue to use questions based on knowing facts and concepts for tests between 2003 and 2011. This pattern is surprising given that in 1997 when TSLN was announced, the then Prime Minister announced that content would be reduced so that there would be more time in schools for project work in order to develop creative thinking and learning skills (C. T. Goh, 1997). Starting in 1997, deliberate and strategic changes were made to some key-stage assessments to reduce the emphasis on content and facts. For instance, in 1999, open-book papers were introduced for A-level Literature, and students were required to devise experiments for A-level Physics and Chemistry (Ministry of Education, 1998). More significantly, project work was introduced as school-based assessment, also at A-level. These changes to this high-stakes assessment are strategic because they signal the types of skills and dispositions that school leavers need to acquire, and hence, should prompt educators teaching the upstream secondary and primary levels to effect similar changes in their curriculum and assessment as well. For instance, the inclusion of performance tasks like project work in a high-stakes assessment at the start of TSLN in 1997 could explain why nearly two-thirds of students had teachers who reported that they placed emphasis on performance tasks as well (see Table 4.3).

There are several reasons why middle school teachers emphasized the assessment of facts and concepts. First, it is necessary to know basic facts and concepts in order to engage in higher-order problem solving. International benchmarking studies like TIMSS include *knowing* as one of the cognitive domains because factual knowledge enables students to engage in more complex cognitive tasks (Mullis, et al., 2009). Second, the increased focus on facts and concepts might be

due to the way teachers view learning. These teachers may possibly focus on facts and concepts because they embrace the view that knowledge has a step-like structure, and students need to master the basics before advancing to more complex tasks and knowledge (Resnick, 1989; Tinzmann, et al., 1991). Such assessment practices resonate with features of behaviorist assessments in that they view learning as sequential and hierarchical (Shepard, 2000). As a result, the middle school teachers focus substantially on assessing basic information before progressing towards more complex knowledge and skills at the higher levels. Further details of such views and assessment practices will be presented in Chapter 6.

The pattern of *change* is illustrated by the statistically significant increased percentage of students taught by teachers who assessed the application of knowledge and understanding between 2007 and 2011. This pattern could be explained by the modifications and reviews of the national science curriculum, given that all Singapore students study content in the centralized curriculum. As mentioned earlier, at A-level, a component of the science practical assessment requires students to design their own experiments (Ministry of Education, 1998). Thus, the changes to the A-level science assessment may be one of the reasons for the small increase in percentage of students whose teachers "sometimes" and "always or almost always" assessed the development of hypotheses and the design of scientific investigations between 2007 and 2011. Another plausible reason for this pattern is curricular change. An analysis of the lower secondary (middle school) science curriculum documents from 1993 to the present day points to the changing cognitive domains and philosophy of science education:

- 1993: Assess knowledge with understanding, handling information; exploration and investigation, and attitudes and development (Curriculum Planning and Development Division, 1992).

- 2001: Assess knowledge with understanding; handling, applying and communicating information, and exploration and investigation (Curriculum Planning and Development Division, 2000).
- 2008: Assess knowledge, understanding and application of science concepts, skills and processes, ethics and attitudes (Curriculum Planning and Development Division, 2007).

As can be seen from the list above, despite the review of the science curriculum, knowledge of science concepts remains a keystone in the assessment. This is aligned with *disciplined inquiry*, one of the AIW criteria. At the same time, the syllabus also emphasizes process skills, and more recently, the objectives even include assessing ethics and attitudes in science. This provides a possible reason for the sustained emphasis on the *applying* and *reasoning* cognitive domains from the period 2003 to 2011. The emphasis of Singapore's national curriculum on these higher-order skills is a plausible reason for the sustained patterns in teachers' responses to the survey items on the assessment of the cognitive domains.

Conclusion

In conclusion, over the 15-year TSLN period, Singapore students were taught by teachers who used a *variety* of assessment practices that included both formal and informal assessment, and adopted a range of objective and constructed-response assessment types. This varied use of assessment practices dominated the early TSLN phase (1995-2003 period). Most of the reported *changes* in assessment patterns occurred in the late phase (2003-2011) in which there were statistically significant differences in the frequency of assessment and in the assessment of higher-order cognitive skills like *application of knowledge and understanding*. The increased

assessment of higher-order cognitive skills may be associated with students' higher scores in the *reasoning* domain in TIMSS 2011.

The hybridization of assessment practices in the early TSLN period resonates with the extant research (e.g., Ohlsen, 2007; Tyack & Cuban, 1995) which suggests that teachers coped with the initial stages of the reform by using a combination of approaches. This is the most efficient way for teachers to adapt to and to incorporate new reforms into their existing repertoire of strategies and their classroom contexts (Ohlsen, 2007). Comparatively, based on the survey responses, changes to assessment practices were reported in the later period, from 2003 to 2011. This is evident by the decrease in frequency of assessing, and the increased emphases on the assessment of the *applying* and *reasoning* cognitive domains.

The shifts in practices in the later phase may be interpreted as responses to structures implemented to realize TSLN. Policies like TLLM further defined and articulated the types of assessment that are aligned to the TSLN vision, as well as reviews of and revisions made to the national syllabuses and assessments. As a result, based on the data reported in this chapter the changes in the assessment practices in the later phase of TSLN appear to have been policy-directed and driven. Unlike short-lived reforms which frequently do not deepen or spread within a system (A. Hargreaves & Goodson, 2006), in the decade and a half period since TSLN was implemented, based on the survey data, the policy seems to have sustained and created change in teachers' assessment practices. Finally, the *persistence* in assessing facts and content continues throughout the entire 15-year period, an indication that there are immutable structures and tools such as high-stakes classroom and school summative assessments that continue to exist in the education system and to impact assessment practices. As a result, teachers continue to stress and emphasize these aspects of assessment. Overall, the combination of *persistence* and *variety* in

the reported assessment patterns indicate “incremental” change (Cuban, 1993) evident in the teachers’ application of hybrid assessment practices while continuing to hold on to or emphasize existing approaches.

CHAPTER 5:

TEACHERS' ASSESSMENT AT THE CLASSROOM LEVEL (2012): QUANTITATIVE ANALYSIS

Introduction

The discussion of teachers' responses to five cycles of TIMSS questionnaires in Chapter 4 provided a *macro* picture of teachers' assessment practices over time. The survey findings reported in Chapter 4 indicated that Singapore students had teachers who used a variety of assessment practices and placed emphasis on assessing higher-order cognitive skills such as applying and understanding. The patterns of assessment practices during the first15 years of TSLN indicated incremental changes as teachers adopted a range of assessment practices while yet maintaining existing approaches. However, these self-report data may over- or under-estimate the frequency and nature of the assessment practices as well as the extent of their use (Mayer, 1999), and they do not provide details of the nature of the assessment practices.

To provide greater insight into the complexities of teachers' assessment practices, eight Singapore teachers were interviewed over a five-month period to examine current patterns of assessment practices used in the course of a school year. During the interview sessions, the teachers submitted assessments, and discussed these in relation to their instructional objectives. They also provided as well as their interpretations of their students' work completed in response to these assessments. Focusing on the *micro* or classroom level brings the analysis of teacher assessment to the heart of the classroom to examine the quality of the tasks, and to examine the extent to which the classroom assessment practices were aligned to the TSLN vision in the fifteenth year of the policy.

Together, both Chapters 5 and 6 present the micro or classroom data in order to illustrate how teachers elicit and enhance student learning through the classroom assessments they use.

Chapter 5 uses quantitative analyses to examine the quality of assessments that teachers use in the classroom and the quality of student work completed in response to these assessments. Chapter 6 presents qualitative analyses of how teachers viewed "assessment," how they practiced formative assessment, and how they identified the conditions that enabled them to use and develop assessments that examined the types of higher-order thinking skills envisioned in TSLN. In short, Chapter 6 provides explanations for the assessment practices used by the eight teachers participating in this study.

The overarching research question examines how Singapore geography teachers have elicited and enhanced student learning through the ways they use classroom assessment since the implementation of TSLN and TLLM. In this dissertation, enhancing student learning refers to teachers' assessment tasks used to elicit student learning as well as the formative assessment practices aimed at closing the gaps and misconceptions in student learning. The findings reported in Chapter 4 provided a national view of the patterns of teacher assessments as well as the quality of student learning over time. The broad picture of assessment tasks used provides an indication of the way the teachers elicited student learning. Students' performance in the content and cognitive domains in TIMSS was used as an indicator of student learning.

In order to o delve more deeply into the nature and quality of assessments teachers use, this chapter presents the analyses of teacher assessment and student work collected over a five-month period. The analyses of teacher assessments using the authentic intellectual work (AIW) criteria (Newmann & Associates, 1996) are a way to examine if teachers are implementing assessments to prepare their students for the *test of life*—that is to say, this dissertation examines if the teachers' assessments assess the types of higher-order skills that will prepare students to contribute to and function in society as envisaged by TSLN. The examination of student work

completed in response to these teachers' assessments is a means by which to elicit the quality of student learning. In this way, Chapter 5 extends the findings presented in Chapter 4. It links the *macro* (national) and *micro* (classroom level) data to examine the quality of teacher assessment and student learning. This chapter looks at the following sub-research questions:

a. What is the nature and quality of classroom assessment that Singapore geography teachers create for their students?
b. What is the nature and quality of work that students produce in response to teachers' classroom assessment?
c. What is the relationship between the nature and quality of teachers' classroom assessment and student work?

Together, the primary and secondary data provide macro and micro patterns of Singapore teachers' classroom assessment practices conducted and implemented in response to the TSLN vision in order to ascertain if teachers were preparing students for the *test of life*. The macro data provided patterns of change over time while the micro data present a picture of current practices.

Background

The selection of the eight teachers for the *micro* part of the study was related to the *macro* aspect of the study. Given the small size of Singapore's education system, it was necessary to include all schools in order to meet the criterion for the minimum student sample size required by the TIMSS sampling guidelines. Therefore, all Singaporean secondary schools were included in each TIMSS cycle (Table 4.2). Because Singaporean secondary teachers would be from schools assessed by TIMSS, as an initial step, all secondary schools listed in the Ministry of Education's School Information Services were contacted using an invitational flyer (Appendix 4).

Teachers participating in the study met one or more of the following criteria: (1) currently teaching lower secondary (Grades 7 and 8) geography; (2) having taught geography for over five

years; and (3) having taught students who had participated in TIMSS 2011. The purpose of Criterion 2 is to gather comments from teachers who have knowledge of the educational policy that forms the context of the study. The eight teachers were selected based on the type of school in which they taught, their meeting the criteria stated in the letter of invitation, and the order in which the responses were returned.

Table 5.1 presents a summary of the eight teachers, their teaching experience and school type. There is a skew towards those who have taught for over a decade. However, this is appropriate as these teachers would then be able to make comparisons between the current situation and that when TSLN began in 1997. Although a group of eight teachers is far too small to be nationally representative, the teachers come from a mix of public, government-aided and independent schools. With the exception of Harry and James, the teachers are all female. Of the eight teachers, Margaret, who is from the European Union, is both the youngest and the only non-Singaporean teacher.

Table 5.1
Participants' background

Participant ID	Years of experience[a]	School Category	Student Population
Rita[b]	22	Senior educator	-
Maryanne	19	Public	Mixed
Harry	15	Public	Mixed
Amanda	15	Public	Mixed
Miki	14	Government-aided	Girls
Totoro	12	Independent	Girls
Jiajia	5	Public	Mixed
James	5	Public	Mixed
Margaret	3	Government-aided	Boys

[a]Arranged in order of years of teaching experience.
[b]With the exception of Jiajia, pseudonyms are self-selected by the teachers.

In addition to the eight teachers from the schools, one more participant, Rita—a senior teacher-leader—was recruited. She has 22 years of experience teaching geography in a number of public schools in Singapore. She played two roles in this study. First, she was one of three raters examining the teacher assessment and pieces of student work. Second, her views were used to triangulate with the researcher's interpretations of the interview data.

While TIMSS assesses four science components (physics, chemistry, biology and earth science) at Grade 8, the micro-level study focused solely on earth science or geography, as this component is called in the Singapore curriculum. First, assessments in geography have been given little coverage in the extant research, both internationally, or in Singapore. Among the studies presented in the literature review in Chapter 2, no more than ten studies focused on assessments in geography. Second, at the lower secondary level, there are parallels between Singapore's geography syllabus and the earth science component in the TIMSS 2011 Science Assessment Frameworks (mapped out in Table 5.2). Third, the nature of geography as a discipline lends itself to the assessment of higher-order-thinking, given the subject's focus on knowledge integration between and within human and physical geography. Fourth, the researcher is a geography specialist with deep interest in and passion for the teaching and learning of the discipline.

Table 5.2
Comparison of Lower Secondary Geography syllabus to the TIMSS 2011 Assessment Framework

TIMSS 2011[a]	Singapore Lower Secondary Geography[b]
Earth's structure, physical characteristics and resources • Sources of water • Existence of air • Feature of the landscape related to human use • Importance of responsible use of resources	Physical environment • Types of landforms • Landforms and people • Distribution of earth's water
Earth's processes, cycles and history • Movement of water • Water cycle • Changes in weather conditions • History of the earth related to fossil remains	Components of the physical environment • Weather and climate • Hydrologic cycle • Weather and climate
Earth in the solar system • Planets and the moon • Sun as source of light and heat • Patterns of the earth's rotation	Earth as home • Earth as part of the solar system • Earth's revolution and rotation
Ecosystems • Impact of human behavior on the environment • Effects of pollution and ways to reduce or prevent	Human environment • Fragile nature of earth • Inter-relationships between people and the environment • Role of humans in managing the environment • Impact of human activities on the environment (protecting and conserving)

[a]Extracted from Mullis, et al. (2009)
[b]Extracted from CPDD (2005).

Nature and quality of teacher assessment

To examine if Singapore teachers are preparing students for the *test of life*, this dissertation uses the *authentic intellectual work* or AIW criteria (Newmann & Associates, 1996) as indicators of higher-order thinking as envisioned by the TSLN vision. The three overarching

AIW criteria are *Construction of Knowledge*, *Disciplined Inquiry* and *Value Beyond School*, and they are further sub-divided into seven standards:

Criterion 1: Construction of Knowledge:

Standard 1: Organization of information
Standard 2: Consideration of alternatives

Criterion 2: Disciplined Inquiry:

Standard 3: Disciplinary content
Standard 4: Disciplinary process
Standard 5: Elaborated written communication

Criterion 3: Value Beyond School:

Standard 6: Problem connected to the world beyond the classroom
Standard 7: Audience beyond the school (Newmann & Associates, 1996, p. 29).

As the AIW criteria emphasize the organization, evaluation, synthesis and interpretation of information, rather than focus on the reproduction, recall, recollection and rote-memorization of learned knowledge and routines, I suggest that the criteria are appropriate to be used as indicators of skills that students need when they leave school. It is also appropriate to use the AIW as a lens to study the teachers' assessments because the authentic achievement goal to "nurture independent, critical thinking in students, and …to help students appreciate, live with, and experience the joy of working with cognitive complex problems" (Newmann & Associates, 1996, p. 44) resonates strongly with the TSLN objective of encouraging Singaporeans to adopt a "spirit of innovation" in order to initiate and drive change to improve society (C. T. Goh, 1997). The AIW criteria also resonate with constructivist assessment, as presented in Chapter 2. Since the AIW criteria focus on some of the higher-order thinking and problem-solving skills that are deemed necessary and useful to individuals and society (Newmann & Archbald, 1992), my assumption is that higher teacher AIW scores mean that teachers are presenting their students

with assessments that focus on the type of higher-order thinking skills aligned with the TSLN vision.

Newmann and Associates (1996) used the seven standards to rate the nature and quality of teacher assessment. Later studies (e.g., King, et al., 2001; Newmann, Lopez, et al., 1998) used a revised version that employed three standards rather than seven for rating the quality of teacher assessments. This dissertation applies the original seven standards to the rating of the teachers' assessments because they are aligned theoretically with the spirit and intent of TSLN. The standards also parallel the assessment objectives and skills in the lower secondary geography syllabus (Table 5.3).

Table 5.3
Comparison of AIW standards and lower secondary geography assessment

Authentic intellectual work standards[a]	Geography assessment objectives and skills[b]
Criterion 1: Construction of Knowledge	Critical understanding and constructing explanations
Standard 1: Organization of information	Select, organize and apply concepts, terms and facts learnt
Standard 2: Consideration of alternatives	
Criterion 2: Disciplined Inquiry	Interpreting and evaluating geographical data
Standard 3: Disciplinary content	Comprehend and extract relevant information from geographical data (numerical, diagrammatic, pictorial and graphical forms)
Standard 4: Disciplinary process	
Standard 5: Elaborated written communication	Use and apply geographical knowledge and understanding to interpret geographical data
	Recognize patterns in geographical data and deduce relationships
Criterion 3: Value Beyond School	Provide holistic understanding of physical-human relationships at the local, regional and global scales
Standard 6: Problem connected to the world beyond the classroom	
Standard 7: Audience beyond the school[c]	Demonstrate a sense of appreciation and responsibility for the quality of the environment at local, regional and global scales
	Demonstrate sensitivity towards people in different human environments

[a] Source: Newmann and Associates (1996, p.29).
[b] Source: Curriculum Planning & Development Division (2005, pp. 1-3).
[c] The syllabus does not have an equivalent for this AIW standard.

There are similarities between MOE's assessment guidelines in the lower secondary geography syllabus and the AIW standards (Table 5.3) which make it appropriate to apply the seven standards in order to rate the teacher assessments. For instance, Singapore's geography syllabus requires that students be able to interpret and evaluate geographical data—disciplinary skills that are necessary in geographical education. This emphasis is similar to the AIW's d*isciplinary process* standard (Standard 4) which rates the extent to which the task requires students to apply inquiry methods characteristic of an academic discipline. In addition, like the

AIW standards which examine the extent to which the tasks present students with issues, questions or problems that students will encounter or are likely to encounter in their lives outside school, the geography syllabus also requires students to apply their understanding of geographical phenomena at local, regional and global environments. Thus, the examination of the teacher assessments using the AIW criteria (Newmann & Associates, 1996) is one means by which to determine the quality of the assessments that teachers present to their students.

Since TSLN's launch in 1997, systemic structures have been put in place to realize the policy. In 2004, the *Teach Less Learn More* (TLLM) movement was introduced to articulate classroom practices that would dovetail with the TSLN intent. In terms of assessment, TLLM encourages teachers to use more "qualitative and formative assessing," and to reduce the emphasis on "set, formulaic answers" (MOE [Bluesky], 2005). The overall aim of education is to prepare students for the *test of life* and not subject them to a *life of tests* (MOE [Bluesky], 2005). MOE has supported these initiatives by providing resources and professional development to scale up teachers' assessment practices (Koh, 2011b). In addition, MOE has revised and reviewed assessment modes used in the non-key and key stage assessments. Since 1997, the national curriculum and national examinations have been revised and updated to reflect TSLN's intended goals (C. T. Goh, 1997; Y. K. Tan, et al., 2008). This is reflected in the 2006 geography syllabus which encourages teachers to use a variety of assessment types and modes, such as oral presentations, portfolio, and fieldwork assignments (CPDD, 2005).[10] Furthermore, as shown in Table 5.4, there is a suggested test blueprint to ensure that teachers assess a range of skills at the midyear and end-of-year semestral examinations. This blueprint recommends the combined use of assessment items, such as objective (e.g., multiple-choice in Section A) and constructed-response (e.g., structured in Section C) questions (CPDD, 2005). The

[10] This is the geography curriculum that was in use during the time of this study.

recommendation that structured questions (Section C) comprise 60% of the entire assessment signals the emphasis placed on students being able to communicate their learning through an extended response rather than their simply providing short one word responses or selecting one fixed response from among a number of options (i.e., multiple-choice).

Table 5.4
Lower secondary geography test blueprint[a]

Section	Item Type	No. of questions to be set	No. of questions to be answered	Weighting (%)
A	Multiple- Choice Questions	15	15	15
B	Map Skills	15	15	25
	Basic Techniques	10	10	
C	Structured Questions	6	4	60

[a] CPDD (2005, p.4).

Similar to the AIW standards, Singapore's geography syllabus focuses on higher-order thinking and downplays the recall of facts and content. This is achieved through the use of Assessment Objectives (AOs). The syllabus stipulates that Knowledge (AO1) is to be assessed in conjunction with Critical Understanding and Constructing Explanation (AO2), and through Interpreting and Evaluating Geographical Data (AO3) (CPDD, 2005, p. 4). The weighting for these dimensions is as follows:

- 50%: AO1—Knowledge + AO2: Critical Understanding and Constructing Explanation
- 50%: AO1—Knowledge + AO3: Interpreting and Evaluating Geographical Data

As can be seen, the AOs state that knowledge should be assessed in relation to higher-order skills such as critical understanding, providing explanations and data interpretation, and should not be assessed as discrete, disjointed, and disconnected bits of information. In this way, the AOs in the geography syllabus resonate with the AIW's *organization of information* standard, in that the emphasis is for students to organize, interpret, explain and evaluate information rather than to retrieve or reproduce isolated pieces of knowledge. The AIW criteria resonate with the

former Prime Minister's view that students need to "create new knowledge" in the 21st century as it is "not enough" to know how to use existing knowledge (C. T. Goh, 1999).

Nature of teacher assessments

The findings of Singapore teachers' classroom assessment reported in Chapter 4 indicated three macro patterns over the period 1995 to 2012. There was *variety* in the practices as Singapore students had teachers who used a hybrid of assessment formats (e.g., objective and constructive-response questions). There was *change* in the later phase of TSLN; more students had teachers who reported assessing higher-order cognitive domains like knowledge application and understanding compared to the early TSLN period. Another change pattern was the reduced frequency of testing between 2003 and 2007 when fewer students had teachers who reported giving a test about once a month. Correspondingly, more students had teachers who reduced testing to a few times a year or less. Third, there was *persistence* as students had teachers who continued to place heavy emphasis on classroom tests and assessing knowledge and concepts in spite of the policy intent of asking teachers to assess higher-order skills. With respect to these trends over the TSLN period, what are the patterns that emerge from the assessments collected for the micro study? Did the classroom assessment used in 2012 mirror the macro patterns which had indicated incremental shifts when examined against the TSLN vision? To what extent were assessments used in the last year of TSLN aligned to the policy vision?

To examine these questions, the eight teachers were asked to contribute a 'culminating' assessment for each of three interview sessions. This assessment task could be something that they typically used or one that was challenging. At least one of the three assessments was submitted before the midyear examination in May and the others were submitted after the June vacation. The last assessment was collected at the end of August. Given the short time frame,

and based on the curriculum plans that the teachers had drawn up, some assessments were not implemented at the end of a unit or units of teaching—these pieces were not culminating tasks. For instance, Jiajia's second assessment was a diagnostic piece to determine how well her students were able to interpret population pyramids.

As this research aimed to collect teacher assessments integral to the teachers' and the schools' curricular plans over a five-month period of the school year, the participants were requested not to design assessments specially for the study. Teachers also were not told how the assessments would be analyzed. This was to reduce the likelihood of participants designing assessments that would match the theoretical framework, and hence inflate the findings.

Table 5.5 presents a summary of the 24 assessments that the eight teachers submitted. The assessment pieces covered a range of topics from the lower secondary geography syllabus, were implemented for a variety of reasons, and used different assessment formats. With the exception of the second assessment that James submitted and the third task Maryanne contributed both of which were taken wholesale from published sources, all the assessments were created by the teachers and their colleagues. Other than Jiajia and Margaret who submitted assessments designed for the classes they teach, most of the assessments were created for an entire cohort of students. The cohort was as large as 13 classes with about 30 students per class. This means that the assessments were not likely customized specifically for any individual student or for a class of students, but were designed and implemented like a mini-standardized test with teachers all marking and scoring the pieces of work from an entire year group.

Table 5.5
Teacher assessment summary

Teacher	Purpose	Topic(s)	Item type
Harry			
Interview 1	Research project	Population, Agriculture, Food, Floods	Open-ended task like a report
Interview 2	Research project	Land reclamation, Water supply	Open-ended task like a report
Interview 3	Common test	Land issues, Water issues, Map reading	Short answers, structured questions
Miki			
Interview 1	Class test	Map reading	Short answers
Interview 2	Class test	Photograph interpretation	Multiple-choice, short answers
Interview 3	Class test	Rivers	Short answers, structured questions
Maryanne			
Interview 1	Common test	Natural vegetation	Multiple-choice, fill in the blanks, short answers
Interview 2 [a]	Common test	Plate tectonics	Multiple-choice, short answers, structured questions
Interview 3[b]	In-class worksheet	Rivers	Fill in the blanks, short answers, sketching
Jiajia			
Interview 1	In-class worksheet	Human impact on the environment	Structured questions (Data response)
Interview 2	In-class worksheet (diagnostic)	Population pyramids	Structured questions (Data response)
Interview 3	In-class worksheet	Map reading	Short answers
Margaret			
Interview 1	Class assignment	Map reading	Short answers
Interview 2	In-class worksheet (diagnostic)	Communication	Structured questions, sketching
Interview 3	In-class assignment	Water issues and Sustainability	Short answers, structured questions
Totoro			
Interview 1	Class assignment	Singapore's water issues,	Structured questions
Interview 2	Research (Fieldwork)	Sustainability	Open-ended task like a report
Interview 3	Common test	High tech agriculture	Short answers, structured questions

Teacher	Purpose	Topic(s)	Item type
		Water issues, Agriculture	
James			
Interview 1	Midyear exam	Population, Urban settlements, Map reading, Agriculture	Multiple-choice, short answers, structured questions
Interview 2	In-class assignment	Water issues	Multiple-choice, structured questions
Interview 3	Common test	Water issues	Short answers, structured questions
Amanda			
Interview 1	Geography level test	Rocks, Rivers, Weather and climate	Multiple-choice, short answers, structured questions
Interview 2	Geography level test	Population, Settlement	Short answers, structured questions
Interview 3	Geography level test	Map reading, Agriculture, Settlement, Population, Transport and Communications, Industrialization	Short answers, structured questions

[a, b] For the second and third tasks, Maryanne submitted differentiated tasks, one task each for the high and middle ability students.

Broadly, the pattern of assessments collected at the classroom level mirror the macro patterns presented in Chapter 4. First, the teachers used a *variety* of assessment types, similar to that suggested in the geography syllabus. The eight teachers submitted one midyear exam paper, three research projects, nine in-class assignments, and 11 common tests and class tests. This pattern suggests that teachers rely on different data sources to elicit evidence of student learning rather than relying solely on performance in the common tests.

Second, the teachers assessed students using a mix of objective and constructive response questions. This again parallels the mélange of assessment practices found in the macro pattern in Chapter 4. Among the 24 assessments collected, nine assessments used a mix of objective and constructive-response assessments, while the remaining 62.5% of assessments required students to construct their responses. The larger number of assessments that used constructed-response questions is consistent with the syllabus emphasis on using questions that require students to construct their responses. The requirement for students to construct their responses is aligned with TSLN's intent to prepare students for life after school. Writing out responses is viewed as more valuable for real-world preparation than having students communicate their learning by selecting a response from multiple-choice options (Newmann & Associates, 1996).

The third emerging pattern is that in 2012 (the year the micro data was collected) teachers engaged in higher frequency of testing as compared to the frequency documented in the macro data (data from TIMSS 2011). The survey responses in Chapter 4 indicated that more students had teachers who assessed a few times a year or less while fewer students had teachers who implemented tests at least once a month. Comparatively, nearly half the assessments (n=11) which teachers submitted for the micro level study were administered as a summative test. On average, over a period of 20 weeks, this works out to one assessment a fortnight, which is a

higher testing frequency than that reported in the macro data. The actual frequency was even higher because during the research period, two teachers—Miki and Maryanne—conducted three retests between them.

Overall, the assessments collected for the micro level study suggest some convergence of assessment practices with the TSLN and TLLM intent. This is evident in the slightly higher emphasis on constructed-response questions which require students to elaborate—sometimes briefly on—a concept rather than simply select a response from multiple-choice options. Another indication of convergence with TSLN is the use of research projects which require students to work independently on a topic, to collect, synthetize and analyze information, and to present the response through extended writing. When compared to existing studies, research projects comprise one-third of the assessments collected for this study. This proportion is higher than that reported in an earlier study by Koh and Luke (2009) in which eight projects (0.5%) were collected from 59 Singapore schools between 2004 and 2005. On the one hand, the fact that project tasks make up a larger proportion of the assessments collected for this study as compared to the extant research suggests that teachers are addressing process and research skills in their assessments, skills that resonate with TLLM tenets. On the other hand, because these extended assessment pieces came from only two of the eight teachers, it suggests that some teachers' assessment practices were more aligned to TSLN than others.

The analysis of the 24 assessments collected also indicates some divergence from the TSLN intent. In particular, teachers continued to administer numerous tests in the classroom. This is seen in the larger number of class or common tests collected for the micro study. More specifically, half the assessments (n=11) collected for this study were used as summative

assessments. Comparatively, in an earlier Singapore study by Koh and Luke (2009), tests made up just 12% of the assessments collected.

In conclusion, the analysis of the types of assessments submitted for this study suggests that in the last year of TSLN, teachers' continued to use a hybrid of assessment practices, similar to that reported in Chapter 4. Some of these assessment practices, such as the use of a variety of assessment types, were closely aligned to the TSLN intent. Other practices such as the majority of the assessments adopting the format used in examinations were less aligned to the TSLN vision. Together, this mixed approach supports the suggestion in Chapter 4 that there has been at best incremental change in the assessment practices. This is because the teachers continue to emphasize assessment practices like frequent formal testing, and, as a result, the assessment patterns do not differ dramatically from those being used at the onset of TSLN in 1997. Chapter 6 provides details to explain these eight teachers' assessment practices.

Quality of teacher assessment

The findings from Chapter 4 indicated that Singaporean students had teachers who reported that they frequently assessed higher-order thinking skills such as explanation and justification. However, the quality of the assessments is unclear because respondents may over- or under-estimate the frequency of use in their survey responses (Mayer, 1999). The survey responses also do not provide details of the relative proportion of each cognitive domain that is assessed in teachers' assessment. While teachers reported that they were assessing the *applying* and *reasoning* domains, the proportion of their assessments that comprise these domains is unknown.

In TIMSS, 65% of the items focus on the *applying* and *reasoning* domains. The weighting for *knowledge* is 35%. In the Singapore geography syllabus, the stipulation is for a

50-50% balance between AO2 and AO3—the equivalent of the *applying* and *reasoning* domains in TIMSS. What is the extent of intellectual challenge, as indicated by the AIW criteria, found in these eight teachers' assessments 15 years after TSLN's implementation? While the geography syllabus only provides a guide for the assessment of higher-order thinking, and schools have curriculum and assessment autonomy at the lower secondary level, it is valuable to use an established research instrument, like the AIW rubric (Newmann & Associates, 1996), to examine the quality of the assessments teachers design and use over the course of a school year.

In this study, the "quality" of assessment refers to the extent to which the eight teachers present their students with tasks that are aligned with the TSLN vision of and emphasis on higher-order skills. Since the AIW standards are used as indicators of higher-order skills, high quality assessments, therefore, are those which are assigned high AIW scores while low quality assessments are those which receive lower scores.

Newmann and Associates (1996) designed the AIW rubric to be applicable to any academic discipline, and for any grade level. The original rubric was designed for mathematics and social studies. For this study, the original AIW rubric for social studies was re-worded for use in geography in consultation with the two raters, both of whom are geography teachers with over 15 years of teaching experience. This will be referred to as the AIW-derived rubric. As this study applied the original seven standards using the terminology adapted for geography, the standards are referred to as the AIW standards. As the criteria are the original ones developed by Newmann and Associates (1996), these are referred to as the AIW criteria and the scores are the AIW scores.

In the original AIW rubric, Newmann, Secada and Wehlage (1995) did not develop full descriptors for all the seven standards. For example, for the *organization of information* standard,

the original rubric only stated three levels: high, moderate, and low. Subsequently, Schroeder, Braden and King (2001) from the Research Institute on Secondary Education Reform (RISER) for Youth with Disabilities further elaborated on the AIW rubric and created descriptors for these three levels (see Table 5.6). To this end, in addition to the original AIW rubric (Newmann & Associates, 1996; Newmann, et al., 1995), the AIW-derived rubric was based on the updated descriptors developed by RISER (2001). Table 5.6 provides an example of Standard 1: *Organization of information* which is rated on a three-point scale. This is an example of a descriptor and rating score based on the revision by RISER (2001). The complete AIW-derived rubric used in this study to rate teacher assessments is in Appendix 5.

Table 5.6
Standard 1 derived-rubric: Organization of information

Standard	Descriptor and rating score
Standard 1: Organization of information The task asks students to organize, synthesize, interpret, explain or evaluate complex information in addressing a concept, problem or issue. Consider the extent to which the task asks the student to organize, interpret, or evaluate complex information, rather than to retrieve or to reproduce isolated fragments of knowledge or to repeatedly apply previously learned routines and procedures. To score high, the task should call for interpretation of nuances of a topic that go deeper than surface exposure or familiarity. When students are asked to gather information for reports that indicates some selectivity and organizing beyond mechanical copying, but are not asked for interpretation, evaluation, or synthesis, give a score of 2.	3 = high The task's dominant expectation is for students to interpret, analyze, synthesize, or evaluate information, rather than merely to reproduce information. [a] 2 = moderate There is some expectation for students to interpret, analyze, synthesize, or evaluate information, rather than merely to reproduce information. 1 = low There is very little or no expectation for students to interpret, analyze, synthesize, or evaluate information. The dominant expectation is that students will merely reproduce information gained by reading, listening, or observing.

[a]This descriptor is adapted from RISER (2001, p. 4)

Table 5.7 presents a summary of the original seven AIW standards and the corresponding rating scores. To provide clarity and to ensure consistency in the use of the AIW-derived rubric, the scoring of the teacher assessment adopted broad guidelines from Koh (2011a) who advises that if a task is comprised of different sections (e.g., multiple-choice, short answer, structured questions), the raters should score the assessments based on the "teacher's apparent dominant or overall expectations" (Koh, 2011a, p. 129). The indicators of overall expectations can be the proportion of time or effort spent on the different sections or the percent of marks assigned to the section. Another broad guideline was to assign the lower score when it was difficult for raters to decide between two scores (Koh, 2011a). This is because a higher score should only be given when there is a persuasive case to be made that a task meets the minimal criteria for the higher score (Koh, 2011a).

Table 5.7
AIW score range

Standard[a]	Score range
Standard 1: Organization of information	1 – 3
Standard 2: Consideration of alternatives	1 – 3
Standard 3: Disciplinary content	1 – 3
Standard 4: Disciplinary process	1 – 3
Standard 5: Elaborated written communication	1 – 4
Standard 6: Problem connected to the world beyond the classroom	1 – 3
Standard 7: Audience beyond the school	1 – 4
Range of possible scores	7 – 23

[a]Adapted from Newmann and Associates (1996, p.29) and Newmann, et al. (1995).

From Table 5.7, five of the seven AIW standards have a score range from 1 to 3 while Standards 5 and 7 have a score range of 1 to 4.[11] The overall AIW score for each piece of teacher

[11] These are the score points from Newmann et al. (1995). The developers did not explain why some standards were rated on a 3-point scale while others were rated on a 4-point scale.

assessment is obtained by totaling the scores from each individual standard to obtain a single cumulative score. At the upper end of the scale, the highest possible score is 23 while at the opposite end, the lowest possible score is 7. The midpoint of the scale is 15.

The rating of the teacher assessments involves assigning a numerical score based on the AIW-derived rubric to each assessment submitted by the teachers. To attach a score to the teacher assessments and student work, qualitative data are converted into ranks and scales through a "quantitizing" process (Miles & Huberman, 1994, p. 42) that enables statistical analysis (Teddlie & Tashakkori, 2003). This enables comparisons to be made between the assessments for each teacher, and among the teachers. Subsequent analysis using *t* tests enabled comparisons of the assessments in relation to each of the AIW criteria.[12]

The rating of the teacher assessments was conducted by three raters over three sessions—June, July and August—immediately after each teacher interview phase. Prior to the first rating session, the three raters had practice sessions to agree on the interpretation of the rubric and the way the teacher assessments and student work should be rated. All three raters assigned AIW scores to the 24 teacher assessments.

Two types of inter-rater reliability statistics—consensus and consistency estimates—were computed to estimate the reliability of the scores.[13] The consensus estimate used to measure the reliability of the ratings is the percent agreement for exact and adjacent scores. For the consensus approach, the inter-rater reliability estimates for the three raters range from 33% to 60% for exact agreement, and from 83% to 88% for exact and adjacent score agreement (Appendix 6).

[12] Although the scores in the rubric are ordinal data (Cassidy, 2009), they have been treated as interval data for analysis in earlier research (e.g., Cassidy, 2009; Gleeson, 2011; Newmann & Associates, 1996; Newmann, Bryk, et al., 2001). Leveraging this, I also treated the rubric scores as interval data for analyses like the *t* test.

[13] The consensus estimates approach is suitable for ordinal data, and in conditions when the different levels of the rating scale represent a linear continuum of the construct (Stemler & Tsai, 2008). Consistency estimates measure whether each rater is consistent in using the categories according to his or her own understanding and definition of the scale (Stemler & Tsai, 2008).

The statistical measure used to compute the consistency estimates is Cronbach's alpha (Stemler & Tsai, 2008). For the rating of teacher assessment, the alpha values were 0.77, 0.92 and 0.82 for Raters 1, 2 and 3 respectively. This indicates high levels of consistency for Raters 2 and 3 and a moderate level of consistency for Rater 1.

Findings. Overall, the quality of the assessments the teachers submitted was low. The mean teacher assessment scores for the three time periods ranged from 13.4 to 14.6. These mean scores were below the scale midpoint of 15. Furthermore, as shown in Table 5.8, on the scale ranging from 7 to 23, only seven tasks had scores above the scale midpoint of 15. This indicates that less than one-third of the tasks submitted by the teachers addressed the higher-order thinking skills indicated in the AIW criteria. As the average mean scores for each teacher ranged from 9.2 to 17.3, this suggests that the quality of the assessments was not at the highest possible level for each of the three time periods. No assessment received a rating that was at the top end of the scale (i.e., 20 to 23 points).

Table 5.8
Authentic intellectual work scores (Teacher assessment)

Teacher ID	Time 1	Time 2	Time 3	Mean (*SD*)[a]	Rank
Harry	18	18	16	17.3 (1.2)	1
Totoro	17	19	15	17.0 (2.0)	2
Miki	15	14	16	15.0 (1.0)	3
Jiajia	16	15	13	14.7 (1.5)	4
James	15	12	12	13.0 (1.7)	5
Amanda	12	12	15	13.0 (1.7)	5
Margaret	14	8	15	12.3 (3.8)	7
Maryanne	10	9	8.5	9.2 (0.8)	8
Mean	14.6	13.4	13.8		
SD	2.6	3.9	2.6		
Midpoint of scale	15	15	15		

[a]Arranged in order of the average mean scores for each teacher.

As only 7 out of 24 assessments had scores above the midpoint of the scale, this suggests that in spite of the stipulation of the Assessment Objectives, the majority of the assessments submitted for this study did not require students to demonstrate higher-order skills. This means that the assessments did not have high expectations for students to analyze and interpret information, demonstrate understanding of geography concepts, use inquiry methods related to geography, communicate their learning through extended writing, and address an issue similar to one in life outside of school—skills that TSLN envisioned as essential for learners.

The overall scores based on the AIW-derived rubric varied among the participating teachers. Some teachers had two or more assessments that scored AIW scores above the midpoint of the scale (e.g., Harry), while other teachers had two or more assessments that scored below the midpoint of the scale (e.g., Maryanne). Furthermore, the scores for the assessment tasks submitted by Harry and Totoro were consistently the highest of the eight teachers. Comparatively, the three assessments submitted by Maryanne were consistently at the lower end of the scale. The other teachers showed some variation in the scores assigned to the assessments which they submitted. For instance, the assessments submitted by James were ranked fourth in the first interview, then fifth and sixth in the second and third interviews respectively. This suggests that some teachers used and designed assessments that were more aligned to TSLN than others.

The assigned AIW scores reflect the nature of the tasks that the teachers submitted. In general, the high scoring assessments required students to interpret and make sense of information and data pertaining to an issue (*Construction of Knowledge* criterion), to contextualize and ground their interpretations and understanding within the academic disciplines (*Disciplined Inquiry* criterion) and to complete a task that is related to an issue or a skill that

students will encounter outside of school (*Value Beyond School* criterion). One example of such an assessment is presented in Figure 5.1. This assessment contributed by Totoro received a score of 19—the highest score assigned by the raters. The reason for this score lies therein that the assessment required students to interpret and evaluate information, to explore and explain the relationships among various kinds of information, to examine different points of views or plausible solutions, to anchor the interpretations using geographical theories and concepts, to engage in geographical research, and to engage in a problem that is related to students' lives and encounters outside of school. These characteristics resonate with the AIW standards.

Figure 5.1
High scoring teacher assessment

INTRODUCTION

Fieldwork helps us to develop important skills such as observations and critical thinking. You will investigate a hydroponic farm in Singapore and understand issues relating to it.

PRE-FIELDWORK

In your pair, read about high-tech farming in your textbook / geography notes and the following information.

[*The text printed for the students to peruse prior to the trip is not presented here.*]

The vegetables produced by XYZ Hydroponic Farm[14] are pesticide-free and grown under soil-less conditions using the Dynamic Root Floating (DRF) hydroponic technique in modular greenhouses. The farm has been supplying hydroponic vegetables since 1991.

Visit the farm's website for more information.

DAY OF FIELDWORK

The attire and items to bring for the fieldwork is as follows:

- School PE t-shirt
- Comfortable shoes
- Fieldwork assignment
- Pen / pencil
- Note pad
- Digital camera
- Small tower / wet wipes
- Insect repellent
- Water
- Poncho / umbrella

ANSWER THE FOLLOWING QUESTIONS.

Location:

1. On a map of Singapore, locate XYZ farm. State the address and identify the Agrotechnology Park that this farm is located in. [3]

Farming system

1. Describe the natural and human inputs at the hydroponic farm. In your answer, you should make reference to land, capital, and labor. [4]

2. Describe the output of the hydroponic farm. [2]

3. Choose 2 of your photographs that best show the cultivation of vegetables using hydroponics. Each photograph must be accompanied by captions [2]

[14] Pseudonym to reduce the possibility of the teacher from being identified.

Decision-making

4. From your dialogue with the management and staff at XYZ Hydroponic Farm, what were the key factors that influenced the decision to adopt hydroponic technology for growing vegetables? These factors may be categorized as follows: [5]

- Economic
- Political
- Technological
- Physical
- Social

5. What are the opportunities and challenges XYZ Farm faces in the next five years? [4]

The task presented in Figure 5.1 required students to investigate a hydroponic farm in Singapore and to understand the complex social and geographical issues relating to its organization. The assessment was completed following a fieldtrip to XYZ farm. It was carefully scaffolded into two phases to help these Secondary 1 students who were studying geography for the first time.[15] In the pre-fieldwork phase, students were tasked to conduct research into the background of the farm from the textbook and the farm's website. At the farm, students had to use their observation and thinking skills to answer a set of questions that Totoro had designed. This set of questions provided a structure for these young Secondary 1 students to help them organize the information that they collected.

With respect to the *Construction of Knowledge* criterion, this assessment received the full score of 6, meaning that it received the full rating score of 3 for each of the two standards, *organization of information* and *consideration of alternatives*. Specifically, this task required students to "organize, synthesize, interpret, explain or evaluate complex information in addressing a concept, problem or issue" (Newmann, et al., 1995, p. 81) as students had to construct a narrative using secondary (e.g., textbook and farm information) and primary (e.g., interview questions) data sources. Second, this task scored the maximum of 3 points on

[15] Geography is only introduced at Secondary 1 as it is not a subject taught at the primary level.

consideration of alternatives because, as stated in the scoring guide, there were no fixed responses for some of the questions (i.e., Questions 4, 5, and 6). For these questions, students could structure their responses based on the way they interpreted the information. In addition, the task allowed students to select a picture that, for them, best captured the hydroponics cultivation process (Question 3).

Totoro's fieldwork task scored 9 out of 10 for the *Disciplined Inquiry* criterion. Each of the three standards—*disciplinary process*, *disciplinary content* and *elaborated written communication*—received rating scores of 3. Elements of *disciplinary content* were evident under the thematic headings of "location," "farming system" and "decision making." For "farming system," for example, the task required students to demonstrate their understanding of the inputs, outputs and cultivation process. The assessment had high expectations for students in terms of *disciplinary process* because fieldwork and interviewing are methods of inquiry central to geographical education. Finally, this assessment scored 3 for the *elaborated written communication* standard because it expected students to provide elaborations and explanations, and to draw conclusions through extended writing.[16] Totoro's task did not score 4 for the *elaborated written communication* standard because the task mirrored a "report / summary" more than "persuasion and theory" as indicated in the rubric.

Totoro's task received a score of 4 out of 7 for the *Value Beyond School* criterion. The main topic in her fieldwork assessment is Agriculture and the complexities in food production, as manifested in the expectation for students to understand the input, output, and cultivation processes. Providing food in land scarce Singapore is a complex issue that students will encounter in life beyond the classroom. Thus, this assessment scored the maximum of 3 on the

[16] Out of a score range of 4, this task received 3 for the *elaborated written communication* standard. According to the rubric, a score of 3 is assigned if the task requires students to write a report or summary while a score of 4 is assigned if the task requires students to adopt a persuasive stance (Newmann, et al., 1995).

problem connected to the world beyond the classroom standard because the difficulties of providing food are very pressing in Singapore. However, the task only scored the minimum score of 1 out of 4 for the *audience beyond the school* standard, since the final product was to be presented only to the teacher.

Overall, Totoro's assessment captured the essence of authentic intellectual work that resonates with the TSLN vision because the task expected students to interpret and evaluate primary (interview) and secondary (textual) data, to integrate relationships among different information sources and variables, to arrive at and explain conclusions, and to discuss responses with more than one plausible solution. Tasks like these which received high AIW scores had features, were aligned to the syllabus. While the format used was different from the suggested test blueprint, the types of skills and content assessed strongly resembled the Assessment Objectives in the syllabus. In particular, the prompts were designed to assess knowledge (AO1) in conjunction with Critical Understanding and Constructing Explanation (AO2) and Interpreting and Evaluating Geographical Data (AO3).

In comparison to Totoro's assessment, more than half of the assessments received scores below the midpoint of the AIW scale. These assessments typically only required students to provide minimal elaboration when communicating their ideas, asked them to reproduce concepts and definitions, and did not require students to engage in any disciplinary inquiry methods. These low-scoring assessments used prompts that only addressed factual knowledge that is unrelated to a real world context or situation, and did not require students to apply any inquiry processes related to the discipline. One example of a teacher's assessment that scored on the lower end of the AIW scale was contributed by Maryanne. This task was implemented following a fieldtrip to a tropical rain forest, and some of the prompts shown in Figure 5.2 below illustrate

clearly Maryanne's expectation for her students to demonstrate their ability to reproduce facts and recall discrete pieces of factual knowledge that can be found in the textbook. These prompts did not ask students to undertake any analysis or evaluation, such as to examine how the features enable the rainforest to adapt to the environment. To this end, assessments that received lower AIW scores did not expect students to demonstrate higher-order skills. As the prompts in these assessments focused solely on assessing knowledge, the tasks are therefore not aligned with the TSLN syllabuses. Such prompts departed from what was stated in the geography syllabus, in particular, that knowledge (AO1) should be assessed with constructing explanation (AO2) and interpreting geographical information (AO3).

Figure 5.2
Example of task focusing on recall of facts

Multiple-choice questions

1. Which of the following is not a major type of natural vegetation?
   a. Grasslands
   b. Tropical rainforest
   c. Corals
   d. Desert vegetation

4 Tropical rainforests are found in an area that has an annual rainfall of _____ mm.
   a. 1000-2000
   b. 550-1500
   c. 2000-4000
   d. 50-300

Short answer questions

2 From what you had gained from your trip to the Botanic Gardens, fill in the blanks below.

Name of plant	Uses
	Flexible, able to bend to make furniture
	Tonic to improve blood circulation

Given the nature of the prompts used in the assessment, it is not surprising that Maryanne's AIW score of 10 out of 23 was on the lower end of the scale. This score is also

below 15, the midpoint of the scale. For the *Construction of Knowledge* criterion, Maryanne received a score of 2 out of 6 because there was little or no requirement for students to analyze or interpret information (*organization of information* standard), and there was no requirement for students to *consider alternatives*. The task also scored low on the *Value Beyond School* criterion because the prompts used did not make references to contexts, issues or problems that students would encounter in their daily lives *(problem connected to world beyond the classroom* standard) and because the final product was only presented to the teacher (*audience beyond the school* standard). For the *Disciplined Inquiry* criterion, the task received moderate AIW scores because it only expected students to demonstrate some understanding of biogeography concepts (*disciplinary content* standard), and annotate diagrams (*disciplinary process* standard*)*. Of the three standards under *Disciplinary Inquiry*, Maryanne's assessment scored the lowest for the third standard, *elaborated written communication*, because the task only required students to complete multiple-choice items and fill-in-the-blank responses. The test format for Maryanne's assessment resembled the test blueprint in that it comprised a mixture of multiple-choice and constructed response questions. However, for the latter, students were only required to provide one word responses, or a phrase. As a result, it is clear that this assessment did not adopt the weighting of higher-order skills as recommended in the Assessment Objectives. Instead of knowledge being assessed with the other assessment objectives, the prompts Maryanne used (Figure 5.2) clearly showed that the focus was on assessing disconnected facts.

In general, the eight teachers addressed some aspects of higher-order skills more than others. As the variations among teachers' scores in the three AIW-derived criteria indicate (Table 5.9), more teachers used assessments that required students to demonstrate academic knowledge than to show their ability to analyze data and apply learning to problems and issues

occurring in the world outside school. Specifically, six out of eight teachers had mean *Disciplined Inquiry* scores that were above the midpoint of the scale. In contrast, three teachers had mean *Construction of Knowledge* scores that were higher than the midpoint on the scale and no teacher had mean *Value Beyond School* scores that were above the midpoint of the scale. The finding that low *Value Beyond School* scores were assigned suggests that across the 24 assessments, teachers did not provide many opportunities for students to apply their learned knowledge to situations that they will encounter in their lives outside of school, nor did they expect students to communicate their learning to a person or persons, other than the teacher.

Table 5.9
Comparison of mean AIW criteria scores

Teacher	Disciplined inquiry Mean (*SD*)[a]	Construction of knowledge Mean (*SD*)	Value beyond school Mean (*SD*)
Totoro	8.7 (0.6)	4.3 (1.5)	4.0 (00.)
Harry	7.7 (0.6)	5.7 (0.6)	4.0 (00.)
Jiajia	7.3 (0.6)	3.7 (1.6)	3.7 (0.6)
Miki	7.3 (0.6)	4.7 (1.5)	3.0 (1.0)
Margaret	5.3 (2.6)	3.7 (1.5)	3.3 (0.6)
Amanda	6.7 (0.6)	3.3 (1.5)	3.0 (0.0)
James	7.3 (1.5)	2.7 (0.6)	3.0 (0.0)
Maryanne	4.4 (1.1)	2.0 (0.0)	2.6 (0.6)
Midpoint	6.5	4	4.5
Max	10	6	7
Min	3	2	2

[a]Arranged in order of mean scores for the *Disciplined Inquiry* criterion.

The eight teachers paid more attention to some cognitive domains than others. As Table 5.10 shows, analyses from a paired *t* test provided statistical evidence that the teachers were more likely to set assessments that focused on *Disciplinary Inquiry* than on *Construction of Knowledge*. Second, the teachers were also more likely to assess *Disciplined Inquiry* than to require students to apply what they had learned to real world problems (*Value Beyond School*

criterion). Last, there was no statistical difference in teachers' emphasis on higher-order skills related to *Construction of Knowledge* and *Value Beyond School*. In particular, the statistical analyses indicate that these eight teachers' dominant expectation was for their students to demonstrate skills related to *Disciplinary Inquiry*. That is, they want their students to demonstrate academic knowledge (*disciplinary concept* standard), use *disciplinary processes*, and communicate this knowledge through extended writing (*elaborated written communication* standard).

Table 5.10
Mean teacher assessment scores by authentic intellectual work criteria (n=24)

Criteria	Mean (*SD*)	Possible score range	Scale Midpoint
Construction of knowledge[a] Organization of information Consideration of alternatives	3.6 (1.5)	2-6 (4)	4
Disciplined inquiry[b] Disciplinary content Disciplinary process Elaborated written communication	6.7 (1.7)	3-10 (6.5)	6.5
Value beyond school[c] Problem connected to the world beyond the classroom Audience beyond the school	3.3 (.67)	2-7 (4.5)	4.5

[a] The results for the paired *t* test indicated that the mean score of 3.6 (*SD*=1.5) for the *Construction of Knowledge* criterion differed significantly from the mean score of 6.7 (*SD*=1.7) for the *Disciplined Inquiry* criterion ($t(26)=10.65$, $p<0.001$). There was no statistical difference in the means between the *Construction of Knowledge* and the *Value Beyond School* criteria ($t(26)=1.3$, $p>0.05$).

[b] The results for the paired *t* test indicated that the mean score of 6.7 (*SD*=1.7) for the *Disciplined Inquiry* criterion ($t(26)=10.65$, $p<0.001$) criterion differed significantly from the mean score of 3.6 (*SD*=1.5) for the *Construction of Knowledge*. The results for the paired *t* test indicated that the mean score of 6.7 (*SD*=1.7) for the *Disciplined Inquiry* criterion was significantly different ($t(26)=11.126$, $p<0.001$) from the mean score of 3.3 (*SD*=0.67) for the *Value Beyond School* criterion.

[c] There was no statistical difference in the means between the *Value Beyond School* and the *Construction of Knowledge* criteria ($t(26)=1.3$, $p>0.05$). The results for the paired *t* test indicated that the mean score of 3.3 (*SD*=0.67) for the *Value Beyond School* criterion was significantly different ($t(26)=11.126$, $p<0.001$) from the mean score of 6.7 (*SD*=1.7) for the *Disciplined Inquiry* criterion.

Overall, the AIW scores indicate that the quality of higher-order assessments used by these teachers was low. When the AIW scores were disaggregated, the ratings assigned to each AIW criterion were also low. As shown in Table 5.10, only the *Disciplined Inquiry* criterion had mean scores above the scale midpoint. Even so, the mean *Disciplined Inquiry* score of 6.7 was not located on the higher end of the scale for this criterion. This suggests that over the five-month study period, the teachers did not frequently expect students to make connections across topics and theories central to the discipline, to engage in disciplinary inquiry, or to communicate their ideas in extended writing. In fact, in many of the tasks, the teachers merely required students to explain or define a concept, meaning that the prompts focused solely on the *disciplinary* content standard under the *Disciplined Inquiry* criterion. For instance, in her second assessment, Margaret asked students to define the term, 'Communication,' and in one task, James tested his students' understanding of "renewable natural resources" and "scarce natural resources." These prompts did not address *disciplinary process* and *extended written communication* because students did not have to apply inquiry methods related to the discipline. While the above prompts requiring students to define or explain concepts spoke to *disciplinary concepts*, the demands made on students were low because

> reference to isolated factual claims, definitions, algorithms—though necessary to inquiry within a discipline—will not be considered indicators of significant disciplinary content unless the task requires students to apply powerful disciplinary ideas that organize and interpret the information (Newmann, et al., 1995, p. 82)

As Newmann et al. (1995) conceptualized the AIW standards, a prompt that has a high demand for the *disciplinary content* standard would require students to integrate one or more geographic concepts, such as this used in Amanda's third task:

> Explain why human activities have a much more powerful impact on Earth now than 50 years ago, in terms of (i) human population, and (ii) transport and communications and industrialization (Assessment 3).

This prompt captured the essence of *disciplinary content* because it focused on key concepts in geography, like change over time, impact of human activity on the environment, and inter-connections across different topics of the syllabus (population, transportation, communications, and industrialization). This is a sophisticated skill demanded of middle school students, given that Amanda's students were introduced to geography just a year previously.

The teachers did not frequently expect students to demonstrate their ability to make meaning from different information and data sources. The *Construction of Knowledge* had a mean score of 3.6, which was below the scale midpoint of 4. This suggests that in the teachers' assessments, there were low to moderate expectations for students to interpret and evaluate data and there was little or no opportunity for students to consider alternative perspectives or points of view as they worked through an issue or a problem. According to the AIW-derived rubric, in a high scoring *Construction of Knowledge* task, teachers would expect their students "to organize, synthesize, interpret, explain or evaluate complex information in addressing a concept, problem or issue" and to consider "alternative solutions, strategies, perspectives and points of view" (Newmann, et al., 1995, p. 81). One example comes from a prompt in the first assessment Totoro contributed (Figure 5.3). Based on the AIW-derived rubric, this assessment scored high on the *organization of information* standard because the nature of the task required students to examine the trends shown in the table, and then suggest possible explanations for the pattern of water use in the hypothetical Country X (*consideration of alternatives* standard). To do well on the task, students could not merely reproduce facts but they needed to draw on what they had learned to support the data trends they were analyzing in relation to the hypothetical context.

Figure 5.3
Example of high-scoring Construction of Knowledge task

3 Study Fig. 3, which shows water usage of Country X, which has a rapidly growing industrial sector. A 'Save Water' campaign was introduced for the first time in 2003.

Year	Annual water usage (million cubic metres)	Population	Water usage per person (cubic metres)
1998	296.7	2 538 066	116.9
2000	300.5	2 546 124	118.0
2002	309.4	2 571 903	120.3
2004	316.9	2 660 789	119.1
2006	320.2	2 718 166	117.8

Fig. 3

Describe and explain the trends on the water usage in Country X.

However, not all tasks had requirements like the one presented in Figure 5.3. As shown in Table 5.10 the mean score for the *Construction of Knowledge* criterion was below the midpoint of the scale. An analysis of the assessment tasks indicates that this is because many of the tasks only addressed one of the two standards under the *Construction of Knowledge* criterion. More specifically, the teachers used prompts that addressed the *organization of information* standard rather than the *consideration of alternatives* standard. To this end, while teachers required students to be able to interpret and synthesize data, they did not provide many opportunities for students to consider "alternatives, solutions, strategies, perspectives, or points of view as they address a concept, problem, or issue" (Newmann, et al., 1995, p. 81), as required in the AIW-derived rubric for a higher-order task. Of the 24 assessments submitted for this study, there were only three instances in which teachers required students to argue from different points of view. For example, in the third assessment she submitted, Margaret posed this question to her students: "Which tap do you think is most sustainable and why?" This question required

students to assess which of the four sources of water supply in Singapore was the most sustainable. For this task, Margaret did not have a fixed response in her mark scheme, although in her view, desalinization would be the least sustainable given its cost and environmental impact. In spite of this, she gave credit to her students as long as "they could justify their opinion." While Margaret's task required her students to provide an "examination of alternatives implicit in the work" (Newmann, et al., 1995, p. 81), this expectation of students was not the general practice. The other teachers in the study, notably James and Miki, only had specific responses in mind, even for discussion questions. For example, although James provided his students with summaries of geographical concepts from a variety of texts, he would only accept responses from the official textbook when preparing responses to his assessments because "they should be reading from the textbook."

As compared to *Disciplined Inquiry,* teachers were less likely to use assessments that required their students to apply their learning to real world contexts and issues as required in the *Value Beyond School* criterion. None of the assessments scored above the midpoint on the scale for this criterion (Table 5.9). The mean score of 3.3 was far below the scale midpoint of 4.5. This suggests that teachers provided few opportunities for students to confront real world issues or problems, and to "communicate their knowledge …for an audience beyond the teacher … and school building" (Newmann & Associates, 1996, p. 29). When examined alongside teachers' emphasis on *Disciplined Inquiry*, teachers seemed to attach more importance to assessing conceptual and factual knowledge than having students relate learning to issues they will encounter when they leave school, or enabling students to communicate their ideas to different audiences.

Summary

On the whole, based on the AIW scores, the nature and quality of assessments submitted for this study are not congruent with the TSLN vision. In particular, the assessments that the eight teachers created over the five-month study period—about 15 years after the launch of TSLN—made low intellectual demands on students and focused on practices that TLLM is exhorting teachers to use *less* of. Fewer than half of the 24 assessments submitted for this study required students to demonstrate the types of higher-order skills envisioned in TSLN as necessary for life outside of school.

Furthermore, when the teachers did address higher-order skills, their focus continued to privilege skills related to *disciplinary content* rather than knowledge construction and application. This suggests a continued emphasis on isolated content knowledge and facts. The teachers were also less likely to require students to interpret and analyze information in relation to problems, issues, and contexts that students are likely to encounter when they leave school. Overall, the types and the nature of the prompts used in these eight teachers' classroom assessments indicate that assessment practices in the later phase of TSLN have not changed much in fifteen years. The pattern emerging from this chapter supports the suggestion in Chapter 4 that the changes in classroom assessment practices are "incremental" (Cuban, 1993). These teachers continued to design and use assessments that modeled the test blue print in its form and format, in spite of the syllabus recommendation that teachers use a variety of assessments (e.g., portfolios, project work) as indicators of student learning. In addition, despite the assessment objectives in the syllabus stipulating a 50% weighting for "critical understanding and constructing explanation" (AO2) and "interpreting and evaluating geographical data" (AO3), the dominant expectation of these teachers' assessments is to test students' ability to reproduce facts and knowledge, rather than to

produce knowledge and apply higher-order skills in their assessments. The interview responses presented in Chapter 6, will provide insight into the ways the teachers elicited and enhanced student learning through their classroom assessment practices, and thereby address why they used assessments that were often not aligned with the TSLN vision.

Nature and quality of student learning

The macro level data presented in Chapter 4 provided evidence of Singapore students' learning over the span of TSLN. First, Singapore students were ranked at or near the top of the scoreboard in science for each of the five TIMSS cycles. Second, the Test-Curriculum Matching Analysis (TCMA) data indicated that despite a decreasing proportion of convergence between Singapore's curriculum and the TIMSS Curriculum framework in each study, Singapore students still outperformed their international counterparts. Third, Singapore students' performance was ranked among the high-performing countries in the TIMSS table in the *applying* and *reasoning* domains in TIMSS 2007 and 2012. In all, Singapore students' repeated strong showing in TIMSS suggests that they are able to respond to test items for which they have not been intensively prepared.

This section uses the AIW criteria to examine the type of higher-order thinking skills that Secondary 2 students demonstrated in response to the assessments set by their teachers in the final year of TSLN. This provides a means of analyzing the quality of student learning at the micro or classroom level during this late phase of TSLN.

Process

The eight teachers were asked to identify 12 students whose work would be submitted for each of the three interviews. As the lower secondary geography syllabus is intended for Secondary 1 and 2, the pieces of student work came from students in both levels. The teachers

submitted completed work from 36 Secondary 1 and 60 Secondary 2 students. Secondary 1 students had been studying geography for about 3 months while Secondary 2 students had been doing so for about a year. In the lower secondary curriculum, geography is taught for about 40 minutes twice a week. A total of 284 assessments from 96 students were collected by the completion of the three interviews. 96 assessments were submitted for Interviews 1 and 2, and 92 assessments were collected during the third interview.[17] To protect the students' identity, the teachers were given code labels for each student.

The rating process was similar to that of the teacher assessments. The rating of student work took into account what students "might be reasonably be expected to do at their respective grade level" (Koh, 2011a, p. 144). Raters scored the assessments independently, and then met to discuss their scores. Prior to the rating session for the first set of student work, about 10% of the pieces were randomly selected and the raters had a pre-session to discuss and to reach a common understanding of the rubric. Two raters rated all the assignments and Rita rated 10% of the pieces of student work for each phase. When there were differences, the raters discussed these with one another in order to reach a consensus. Using the consensus estimates, the percent of exact agreement between two raters ranged from 46% to 68% while the percent of exact and adjacent-score agreement ranged from 84% to 100% (Appendix 7). These values indicate high levels of inter-rater reliability.[18] In terms of the consistency estimates, the Cronbach's alpha values were 0.9 for Raters 2 and 3, and 0.8 for Rater 1. These values are consistent with that reported in previous research.

[17] Three students were on medical leave and one student was representing the school in an international student conference. When computing the data, the scores of the missing students were replaced by the class mean score.
[18] For rating of student work in one of their two data sets, King, et al. (2001) reported that inter-rater percentages of 47.1% for exact agreement and up to 88.4% for exact and adjacent score agreement.

The rating of student work comprises two AIW criteria, *Disciplined Inquiry* and *Construction of Knowledge* (Table 5.11). The *analysis* standard is part of the C*onstruction of Knowledge* criterion. The *disciplinary concepts* and *elaborated written communication* standards are from the *Disciplined Inquiry* criterion. Due to project resource limitations, Newmann et al. (1995) did not create a rating for the *Value Beyond School* criterion when they developed the rubric. The rubric used for rating student work was drawn from the pre-existing one developed by RISER (2001), which in turn was based on the original rubric developed by Newmann and Associates (1996). The adapted rubric used for this study is referred to as the AIW-derived rubric as the descriptors were developed by RISER (2001) . As the standards and criteria are from Newmann and Associates (1996), they are referred to as the AIW standards and AIW criteria respectively. The scores are termed the AIW scores.

Table 5.11
AIW-derived rubric for student work [a]

Standard	Descriptor	Possible score range
Construction of Knowledge Criterion		
Analysis	Student performance demonstrates higher-order thinking in geography content by organizing, synthesizing, interpreting, evaluating, and hypothesizing to produce comparisons, contrasts, arguments, application of information to new contexts, and consideration of different ideas or points of view.	1-4
Disciplined Inquiry Criterion		
Disciplinary concepts	Student performance demonstrates an understanding of geographical ideas, concepts, theories, and principles, and uses them to interpret and explain specific, concrete phenomena, information or events.	1-4
Elaborated written communication	Student performance demonstrates an elaborated account that is clear, coherent, and provides richness in details, qualifications and argument. The standard could be met by elaborated consideration of alternative points of view.	1-4

[a] Adapted from RISER (2001).

Each of the three standards was rated on a four-point scale (Table 5.11). As such, the possible score range for student work is from 3 to 12. The midpoint of the scale is 6.5. For example, for *analysis*, the four points range from 4=substantial evidence of analysis to 1=no evidence of analysis. Table 5.11 presents a summary of the three standards. The details of the student work rubric are in Appendix 8.

Quality of student learning

Singapore students' repeated success in consecutive TIMSS cycles suggests that when presented with intellectually demanding tasks, they were able to produce more complex intellectual responses, even though they were not prepared or drilled to answer the TIMSS achievement items. The extant research has reported that the quality of student work varies with the level of expectations indicated in the assessments teachers use (Bryk, et al., 2000; Newmann & Associates, 1996; Newmann, Bryk, et al., 2001; Newmann, Lopez, et al., 1998). Other studies have reported a direct relationship between the quality of tasks teachers set and student work (Clare & Aschbacher, 2001; Gleeson, 2011; Ladwig, et al., 2007; Matsumura & Pascal, 2003). This means that when teachers use challenging prompts, students are able to provide high quality responses. As a result, the nature and quality of tasks teachers assign become "constraints and thresholds" (Koh & Luke, 2009, p. 312) to the types of responses students produce. Based on the findings reported in the extant research, and given the types of assessments these teachers used, the quality of student work produced in response to these teachers' assessments in this study is expected to be low.

As shown in Table 5.12, the quality of student responses produced in response to the teachers' assessment was indeed low. This means that the quality of learning exhibited in the students' work was not aligned with the TSLN vision of thinking citizens who can contribute to a

learning nation. Only five sets of student work had means above the scale midpoint of 7.5. The overall mean AIW class score for the 284 pieces of student work collected for this study was 6.4 (*SD*=2.3) out of a maximum score of 12. This means that more than half of the sets of student work did not demonstrate a moderate level of higher-order thinking skills. Lower AIW scores indicate that student work demonstrated little analysis, did not apply geographical concepts, and provided minimal elaboration of ideas.

Table 5.12
Authentic intellectual work scores (student work)

Teacher	No. of pieces	Class mean 1	Class mean 2	Class mean 3	Overall mean (*SD*)[a]
Totoro	36	8.9	10.5	7.3	8.9 (1.6)
Amanda	36	7.3	7.3	8.2	7.6 (1.9)
James	36	6.7	7.6	5.3	6.5 (0.9)
Miki	35	4.6	8.8	6.2	6.5 (1.0)
Margaret	36	6.5	4.4	6.9	5.9 (1.2)
Jiajia	36	6.1	6.2	5.3	5.8 (1.1)
Harry	35	6.6	5.2	5.3	5.7 (0.8)
Maryanne	34	5.7	3.4	4.5	4.5 (1.4)
Total	284				
Mean (*SD*)		6.5 (2.2)	6.7 (2.6)	6.1 (2.0)	
Midpoint		7.5	7.5	7.5	

[a] Arranged in order of the overall mean student work scores for each teacher.

The students demonstrated higher-quality work in some AIW standards than others. This is in response to the types of higher-order skills their teachers addressed in the assessment tasks. As shown in Table 5.10, teachers were most likely to use prompts related to *Disciplined Inquiry*. Correspondingly, students' work also showed higher AIW scores for the *disciplinary concepts* standard. As shown in Table 5.13, the results from the paired *t* tests indicated that students were better able to demonstrate understanding of geography content (*Disciplinary concept*) than they

were at analyzing and communicating their ideas clearly and cogently (*Elaborated Written Communication*).

Table 5.13
Mean student work scores by authentic intellectual work standards (n=288)

Criteria	Mean (*SD*)	Possible score range
Construction of knowledge criterion Standard 1: Analysis[a]	2.2 (0.9)	1-4
Disciplined inquiry criterion Standard 2: Disciplinary concepts[b]	2.3 (0.9)	1-4
Standard 3: Elaborated written communication[c]	2.0 (0.8)	1-4

[a] The paired *t* test indicated that the mean score for the *analysis* standard is statistically different from the mean scores for the *disciplinary concepts* standard ($t(287)=3.601$, $p<0.001$) and *elaborated written communication* standard ($t(287)=4.212$, $p<0.001$).

[b] The paired *t* test indicated that the mean score for the *disciplinary concepts* standard is statistically different from the *analysis* standard ($t(287)=3.601$, $p<0.001$) and the *elaborated written communication* standard ($t(287)=6.884$, $p<0.001$).

[c] The paired *t* test indicated that the mean score for the *elaborated written communication* standard is statistically different from the *analysis* standard ($t(287)=4.212$, $p<0.001$) and the *disciplinary concepts* standard ($t(288)=6.884$, $p<0.001$).

Although the statistical analyses indicate that students' work completed in response to the assessments the eight teachers used was strongest in *disciplinary concepts*, the mean score for this AIW standard was low. The score of 2.3 was just below the scale midpoint of 2.5. This score means that students were able to utilize geography concepts in their work "but their use is significantly limited" (RISER, 2001, p. 27). Two student responses to the same question illustrate the difference between a high and a low score for the *disciplinary concepts* standard (Figure 5.4). In this assessment, students were presented with a diagram of the long profile of a river, and asked to select from one of two locations, a site that was feasible for a hydroelectric dam.

Figure 5.4
Comparing student responses for the disciplinary concepts standard

Prompt: Locations A and B have been selected for the construction of a dam to generate electricity. Which of these two locations is a better choice? Justify your answer.	
High Scoring Student (Score = 4)	**Low Scoring Student (Score = 1)**
Location B. There are more tributaries at Position B as compared to A which will help to provide more water. At Position B, there is more water as Position A is near the source of the river where little water can be collected as compared to Position B. It is also very dangerous for the workers to construct a dam at position A because the slope is too steep while the slope at B is less steep. Thus Position B is better.	Location B is a better choice. Location A is the river source, thus water has just been collected and there would not be a lot of water. At Location B, more water is collected, thus if the dam is built there, it can generate more electricity, making it a better choice.

While both students selected the right location—Position B, the high-scoring student's response, demonstrated understanding of hydrological concepts because it explained that at Position B, tributaries flowing into the main river would add to the volume of flow, thus collecting more water to run the turbines of the dam. This student also understood the concept of "gradient" at the upper course of the river which makes it difficult to construct a dam. In other words, the high-scoring student used "exemplary understanding" of geographical concepts to "organize [and] explain … otherwise discrete pieces of information (RISER, 2001, p.27). The low-scoring student applied and used disciplinary concepts in a less precise and sophisticated way, as, for instance, in her use of vernacular language rather than geographical and hydrological concepts.

The second highest mean AIW class score was assigned to responses on the *analysis* standard. The mean of 2.2 fell below the midpoint of the scale, suggesting that student work exhibited "some evidence of analysis" (RISER, 2001, p. 27). This score indicates that students' responses demonstrated more knowledge reproduction than information synthesis or integration. Two students' responses to an assessment prompt, "discuss the differences between rural and

urban settlements," illustrate differences in the quality of work on the *analysis* standard (Figure 5.5). While the responses indicate that both students understood the difference between rural and urban settlements, the low-scoring student's response did not include any analysis. In fact, the response reads like a grocery list of facts about rural settlements followed by another list of discrete details of urban settlements. This suggests that the student merely reproduced the facts as listed in the textbook. Conversely, the high-scoring student's response provided "moderate evidence of analysis" and organization (RISER, 2001, p. 27) to highlight the differences between urban and rural settlements in terms of population size, type of economic activities, social life, and level of provision of amenities. In the high scoring student's work, the characteristics of rural and urban settlements were also peppered with examples.

Figure 5.5
Comparing student responses for the analysis standard

Prompt: Discuss the differences between rural and urban settlements.

High Scoring Student (Score = 3)	**Low Scoring Student (Score = 1)**
A rural settlement has low population size and densities while urban places have high population size and densities. The main function in rural settlements is fishing, mining, and farming while in urban settlements, people do business and manufacturing. In rural settlements, people only meet and communicate with people in their own village, while in urban settlements, people live and work closely together. In rural settlements, there are few amenities like schools, post offices and roads to cater to the needs of the people. But in urban settlements, there is a wide range of amenities for business and industries to run smoothly. There are airports, seaports and specialized medical and banking services.	In rural settlement, there is low population density with not much amenities. They farm, mine, and fish to meet their basic needs. There are only basic amenities such as schools and post offices. In urban settlements, there is a high population density with many buildings close together. People live in buildings which are closely apart. They often do manufacturing and business. There are many amenities.

The *elaborated written communication* standard had the lowest mean score of 2.0. The score was also below the midpoint of the scale. This suggests that students were least able to produce responses that "provide richness in details, qualifications and argument" (RISER 2011, p 27), and even in many cases, not able to demonstrate "reasonably accurate elaboration for at least one important statement" (RISER, 2001, p. 28). Two more student work examples illustrate the different quality of student responses under the *elaborated communication* standard (Figure 5.6). The task required both students to identify and justify the most sustainable source of water supply for Singapore. While both students could identify a particular source tap as a means of sustainable water supply, the difference in the responses is in the way the students elaborated on their choices. The high-scoring student identified and justified why "NEWater" was his sustainable water supply choice after discussing the disadvantages of the alternatives. This response provided "some elaboration for two or three important statements … Arguments or explanations are present. They are concise, clear, and well-articulated" (RISER, 2001, p.28). On the other hand, the low-scoring student only highlighted the strengths of his chosen water source through "discrete claims, [and] broad generalizations" (RISER, 2001, p.28).

Figure 5.6
Comparing student responses for the elaborated written communication standard

To ensure a sustainable water supply, Singapore has put in place a system called the four national taps. (a) Name the taps. (b) Which tap do you think is most sustainable and why?	
High Scoring Student (Score = 3)	**Low-Scoring Student (Score = 1)**
(a) Imported water, desalinated water, NEWater, Reservoirs	(a) NEWater, imported water, local catchment, desalination
(b) NEWater. Firstly, imported water is not a permanent water supply, once the contract expires in the future. Desalinated water is a good way but it is extremely expensive, and therefore not a reliable source. Though Singapore has reservoirs, it is also not a permanent source, since there are too few of them. NEWater, is more affordable, and since it reclaims used water, it would therefore be the most sustainable and reliable.	(b) Local catchment water. A lot of Singapore's water is supplied from local catchments. Local catchment water also has a large water supply that will help to supply Singapore with a lot of water.

It is not surprising that the mean student work score for the *elaborated written communication* standard was low, given that several teachers did not provide students with the opportunity to undertake extended writing of one or more paragraphs. While a large proportion of some assessments included constructed-response questions, they mostly required short responses comprising 2-3 sentences, or short phrases. For instance, while Jiajia required her students to interpret data from charts and tables, the prompts she used, such as "identify the name of the country with the largest urban population" (Assessment 1), only required one word or short responses. Although Jiajia's tasks scored high on *analysis* because of the requirement to interpret charts, data and map symbols, the answers required mostly one or two words, or short responses of between two and four sentences. As a result, students did not have the opportunity to demonstrate their *elaborated written communication* skills. Likewise, in the three assessments

Miki provided for her students, only short responses were required. In the bulk of Miki's first assessment, only one word responses were required. In the second assessment, Miki used a mix of multiple-choice questions and short response questions requiring a sentence or two. And in the third assessment, Miki only required her students to submit short responses requiring a few sentences. To this end, such assessments provide little opportunity for students to demonstrate their learning through extended communication.

Overall, the quality of student work produced in response to teachers' assessments was low. Yet, in international studies such as TIMSS 2011, Singapore students have received high scores for the content and cognitive domains despite just 68% of the international items are aligned to the Singapore Secondary 2 science curriculum in 2011 (see Martin, et al., 2012). This achievement suggests that with strong basics in science content, Singapore students demonstrate that they are able to apply knowledge learned to respond to the questions requiring application and reasoning skills. It therefore follows that when given the opportunity, students are able to demonstrate their competencies in the higher-order domains if the tasks are structured appropriately, as in the fieldwork assignment designed by Totoro.

The extant research reports that there is a direct relationship between the level of expectations in the tasks teachers use and student work (Clare & Aschbacher, 2001; Gleeson, 2011; Ladwig, et al., 2007; Matsumura & Pascal, 2003). This indicates that the nature and quality of tasks teachers use impose a ceiling effect on the quality of work that students produce (Koh & Luke, 2009). To this end, the quality of student work is contingent on the types of assessments with which they are presented. When an assessment makes low demands for authentic work, students will most likely also score low on the AIW standards because they will have virtually no chance to exhibit their proficiency in constructing knowledge or in disciplined

inquiry (Newmann, Lopez, et al., 1998). Comparatively, if teachers have high expectations and create prompts that require students to demonstrate authentic work, there will be opportunities for students to show what they can do (Newmann, Lopez, et al., 1998). Therefore, when teachers like Totoro present tasks that require independent research to organize, synthesis, and analyze geographical issues and to construct a lengthy response, correspondingly, their students work demonstrate higher levels of thinking skills. In comparison, when teachers like Maryanne only use prompts that assessed factual and procedural knowledge, do not require analysis or evaluation, and did not require students to tackle multiple-choice prompts, the resultant student work only shows students' ability to select multiple-choice responses and to recall facts.

Like the extant research, this study also found a direct relationship between teacher assessment and student work. The Pearson Product-moment correlation coefficient was positive and significant (r=0.53, p<0.01),[19] suggesting a direct relationship between teacher assessment and student work. This statistic can be interpreted to suggest that when teachers design assessments that demand higher-order skills, correspondingly, the responses produced by students will exhibit higher-order skills. Conversely, students who receive less challenging work are not expected to produce any authentic work (Newmann, Lopez, et al., 1998). The positive relationship indicates that teachers' assessment practices influence "what will count" as good quality student work (Koh & Luke, 2009) and also, more bluntly, "what you test is what you get" (Koh, et al., 2006, p. 6).

Summary

The analyses of student work completed in response to their teachers' assessment indicate that students are not demonstrating the types of skills necessary for life outside school. This is

[19] Typically when interpreting the Pearson Product-moment correlation, the closer the correlation coefficient to plus and minus 1, the stronger the correlation (Shavelson, 1996).

based on the low AIW scores which indicate that student work included little analysis, demonstrated limited use of geographical concepts, and provided responses with minimal elaboration. Furthermore, student work was also of low quality for each of the different skills embodied in the AIW criteria. Specifically, student work only showed some evidence of their being able to organize and make sense of complex information, students were limited in using and applying disciplinary concepts, and they provided few details and qualifications in their writing.

At the same time, there were examples of student work which demonstrated high levels of competency in the AIW standards, as shown in Table 5.12. These sets of student work were completed in response to teachers' assessments that exacted high cognitive demands on students. As the correlational analyses indicates, the quality of work that students produce is associated with the types of assessment prompts their teachers present them. Since the quality of higher-order prompts as indicated by the AIW criteria used by the teachers in this study is low, it is therefore not surprising that the corresponding student work completed in response to these assessments was indeed low. Based on this analysis, in order for students to demonstrate higher-order skills needed for life outside of school, there is a need for teachers to engage their students in a wider range of higher-order tasks.

Discussion and conclusion

The goal of education in Singapore following the introduction of TSLN was to develop students into thinking citizens, ones who can produce knowledge—a necessary skill needed to survive in a knowledge-based society (C. T. Goh, 1999). To this end, in the past decade and a half, a variety of policies and initiatives were implemented to realize this education vision. The policies—to name a few—included freeing up space in the curriculum for teachers and students

to engage in deep learning (Shanmugaratnam, 2004, 2005b), making changes to the national examinations (Y. K. Tan, et al., 2008), and reminding teachers to re-examine *why*, *how*, and *what* they teach through TLLM ((MOE [Bluesky], 2005).

In Chapter 4, the analyses of the survey responses indicated that three patterns of Singapore teachers' assessment practices emerged over the TSLN period: *change*, *variety* and *persistence*. These three patterns indicated "incremental" change (Cuban, 1993) in which teachers combined new and old practices that enabled them to do their job more efficiently (Tyack & Cuban, 1995) and to enable them to manage the change resulting from the policy (Ohlsen, 2007) as they encountered and worked within the TSLN vision from its inception. While the responses of Singapore students' teachers reported in the TIMSS questionnaires provide patterns of practices at five different time points, the analyses of the 24 assessments collected from the teachers participating in this study provide a glimpse into classroom practices over a 5-month period during the school year. Based on the assessments that eight teachers submitted in the later phase of TSLN, how did teachers elicit student learning through their classroom assessment in the later years of TSLN? To what extent were the assessments used aligned with the TSLN vision?

Overall, three patterns of assessment practices emerge from the analyses of the 24 tasks: (1) the *variety* of assessments used; (2) the *persistence* in some assessment practices; and (3) the *low quality* of the assessments. First, there was is *variety* in terms of types of assessments teachers used. The teachers used a range of assessment types and formats. Of the 24 assessments submitted, there were extended projects, in-class assignments, worksheets, open-book tasks, and tests and examinations. Some teachers submitted summative tests which provided grades used in computing students' final year grade. Other teachers submitted a mixture of in-class worksheets

and common tests. In terms of the assessment format, the teachers used a mixture of constructed-response and objective question types, a pattern similar to that reported in the macro data presented in Chapter 4.

Second, the types of assessments submitted suggested *persistence* in some assessment practices. In spite of TLLM's call to reduce emphasis on "one-size-fits-all" and to increase the use of "differentiated" instruction (MOE [Bluesky], 2005), the assessments—whether formal or informal—continued to be designed as standardized pieces. There was little evidence that the teachers heeded the call for differentiation. None of the assessments were customized to meet the profiles or needs of different students within the class. In addition, despite the autonomy granted to schools, teachers continued to adopt the assessment format recommended in the syllabus for the end-of-the-year and midyear assessments. This means that their assessments mirrored the suggested test blueprint mentioned in the syllabus. Thus, there was little variation in the format of the assessments submitted by most of the teachers. Only two teachers—Totoro and Harry—submitted assessments that had more variation in form and format than their colleagues. There was also continued emphasis on the assessment of facts and concepts, with many tasks comprising prompts that required students to "define" and "explain" concepts and facts, as indicated by the higher mean scores for the *Disciplined Inquiry* criterion.

This pattern indicates that fifteen years after TSLN's introduction, assessment practices were not closely aligned to the policy intent, and teachers continued to mimic the format of the end-of-the-year assessments, rather than to create and use a greater range of assessments. The finding that half the assessments submitted over the study period were comprised of class tests and examinations signals a continued high frequency of conducting a test. The implementation of 11 tests over 20 weeks also indicates a testing frequency averaging one test every fortnight.

The large number of tests and examinations collected indicates a continued emphasis on summative assessments, and suggests an ongoing emphasis attached to assessing the *product* of learning. Only two teachers submitted extended projects, indicating that during the research period, the rest of the teachers did not provide many opportunities for students to address the *process* of learning, an emphasis that TLLM is encouraging teachers to do more of. Finally, as the format used in the common tests adhered closely to the test blueprint for examinations as suggested in the syllabus, it seems clear that over the course of the school year, these eight teachers spent much time exposing their students to the format to be used in the end-of-the-year assessment. Once again, this points to assessment practices that focus on preparing students for a *life of tests*, as the many common tests are indicative of many opportunities provided for students to practice on the assessment formats used for the end-of-year assessment—even though lower secondary is a non-key stage level.

The third pattern that emerges from the analyses of the assessments was the *low quality* of the assessments. The analyses indicated that the teachers did not frequently use tasks that required students to demonstrate higher-order skills as envisioned in the TSLN vision. Overall, less than one-third of the 24 assessments scored above the midpoint of the AIW scale. This indicates that the teachers' dominant expectations in the tasks did not require students to demonstrate skills related to the seven AIW standards. On the contrary, the prompts used mostly required students to reproduce knowledge in the same form as it was learned, to communicate their learning by selecting an option or providing short responses, to be able to produce one correct response that the teacher had predetermined, to know unrelated pieces of knowledge, and to complete tasks that bore little or no resemblance to issues and problems that students would encounter beyond school. Furthermore, the teachers did not frequently assess higher-order

thinking, given that over a span of five months, only 7 out of 24 assessments focused on it. The analyses of the assessments suggest that 15 years after TSLN, the quality of the assessments used in these eight teachers' classrooms does not align with the overall policy objectives of equipping students with the types of higher-order skills needed when students graduate from school. Focusing on cognitive skills as embodied in the AIW criteria is important because the ability to apply and construct knowledge is necessary to thrive in contemporary society where employees from all backgrounds are expected to be able to apply knowledge to solve problems rather than to mechanically and routinely churn out previously practiced procedures and learned content (Bryk, et al., 2000).

As mentioned earlier, to realize TSLN, changes were made to Singapore's curriculum, such as revising, reviewing, and reducing the content so that there is time for the development of higher-order skills. The syllabuses stipulated the use of Assessment Objectives to ensure that knowledge is not tested in isolation, but in conjunction with skills like constructing explanation, and interpreting and analyzing data. However, as the analyses of the 24 assessments indicate, the *quality* of the assessment tasks, based on each AIW criterion was also low. In particular, with the exception of the *Disciplined Inquiry* criterion, the mean AIW scores for the *Construction of Knowledge* and *Value Beyond School* criteria were below the midpoint of the scale. Overall, the AIW scores indicate that the assessments did not contain prompts or tasks that required students to demonstrate higher-order skills such as making sense of and analyzing information and applying learning to real world contexts.

The assessment tasks did not address higher-order skills equally. The teachers addressed some higher-order skills more than others. Specifically, their tasks focused more on skills related to *Disciplined Inquiry* than to *Construction of Knowledge* and *Value Beyond School*.

This means they continued to check students' learning of content and facts as compared to assessing students' ability to apply and transfer knowledge to problems and issues relevant to the world outside of school. The prompts used in the assessments also suggest that the teachers' dominant expectation was for students to reproduce material in the same form in which it had been learned (Newmann, Bryk, et al., 2001). Since very few prompts received scores for the *considering alternatives* standard, it is clear that the teachers only expected specific responses as they had been taught to students, and did not expect to engage students in exploring plausible alternatives, perspectives, and arguments. Thus, the mean scores of the individual AIW criteria indicate that the teachers' assessments were not aligned with the assessment objectives stipulated in the geography syllabus, one that was developed and introduced to realize TSLN. This is because the teachers used assessments that emphasized content knowledge rather than their designing tasks that assessed knowledge in conjunction with "Critical Understanding and Constructing Explanation" and "Interpreting and Evaluating Geographical Data" (CPDD, 2005, p. 4). Furthermore, the higher mean scores for *Disciplined Inquiry* over *Construction of Knowledge* and *Value Beyond School* indicate that teachers paid more attention to assessing students' mastery of disciplinary knowledge. Even so, the *Disciplined Inquiry* means were not at the highest possible level, and this shows that teachers expected students to recall knowledge and facts, rather than to demonstrate their understanding of facts and knowledge in a meta-cognitive sense such as being able to integrate ideas and concepts across topics.

The low AIW ratings found in this study, whether these are the composite scores or individual criterion scores, corroborate earlier research on teachers' classroom assessment (e.g., Koh, et al., 2005; Koh & Luke, 2009) and the "glaring absence of intellectual demand" in the types of assessment set in Queensland (Lingard, et al., 2006, p. 10). Some questions emerge from

these patterns of findings. If TLLM had been initiated to articulate strategies to guide teachers in teaching and assessment, if the syllabuses had been revised to provide curricular and assessment guidance to enable schools to realize the TSLN vision, and if the system has been guided by TSLN for over a decade, why then did the AIW scores remain low? How do teachers perceive assessment issues and what influences them in designing and using assessments that assess higher-order skills? In Chapter 6, the data from the in-depth interviews with the eight teachers will provide some responses to these questions.

The above patterns indicate that the nature of the assessments these eight lower secondary teachers used to elicit student learning diverge from the TSLN vision. The syllabuses designed to realize TSLN allow for autonomy at the school-level, and encourage teachers to use a variety of assessment types to elicit student learning. Yet, with immense fidelity, 50% of the assessments adopted the format suggested for the end-of-the-year examination. This suggests a deliberate and concerted focus in the course of the school year on preparing students for the year end examinations, rather than equipping students with the skills important for the *test of life* as envisioned in TLLM.

The types of learning that teachers elicited in their assessment tasks had implications for student learning. Prior research has reported that when teachers assign more intellectually demanding assessments, their students are able to produce work that reflects higher-order intellectual performance (Bryk, et al., 2000; Koh & Luke, 2009; Newmann & Associates, 1996). To this end, the quality of the assessments these eight teachers presented placed a cap on the types of higher-order skills their students were able to produce. It was therefore not surprising that student work completed in response to these 24 assessments paralleled the level of intellectual challenges that their teachers provided. Based on the assessments presented to them,

the students taught by the eight teachers participating in this study produced work that did not reflect the quality of learning envisioned in TSLN. The overall quality of student work was low. In-depth analyses of the AIW scores indicated that students produced higher-quality responses on some criteria more than on others. More specifically, students demonstrated their ability to handle subject knowledge, as evidenced by their high scores for the *disciplinary concepts* standard. Students did not demonstrate much interpretation or synthesis in their work, based on the mean scores for the *analysis* standard. They were unable to exhibit their ability in written communication—a necessary 21st century skill—as their work scored lowest for the *elaborated written communication* standard.

The students taught by the eight teachers did not always have the chance to exhibit work that indicated mastery of higher-order skills because the assessments they were presented with did not demand it (Bol & Strage, 1996). To this end, if teachers assessed higher-order thinking skills, then their students' work would have exhibited more advanced understanding. Conversely, if teachers implemented tests that only sought to assess recall and rote learning, then students would only have been able to demonstrate these skills. The extant research has reported "encouraging" findings that when teachers do assign their students more demanding tasks, the students' performance indicates that they are able to complete complex cognitive tasks (Bryk, et al., 2000, p. 2). Thus, the more teachers expect from their students, the higher their students will perform (Newmann, et al., 2007). In other words, in order to prepare students for the *test of life*, teachers need to present their students more frequently with opportunities to demonstrate higher-order skills. As reflected in the student achievement data in TIMSS, despite the Singapore science curriculum being about 70% aligned with the TIMSS framework, Singapore students have excelled at the cognitive domains of *applying* and *reasoning* as well as in the content

domains. As students cannot be drilled for the TIMSS achievement items, their performance in this international benchmarking study indicates that when presented with challenging prompts, the corresponding student responses show that the learners are able to produce high-quality work.

Drawing from this analysis, one can infer that the quality of classroom assessment plays a significant role in steering and driving student learning beyond the current assessment of factual knowledge and routine procedures. The analysis of tasks such as Totoro's fieldwork assessment suggests that challenging tasks are able to powerfully engage students. Totoro's fieldwork assessment received the highest AIW score of the 24 assessments. Correspondingly, her students' responses to this assessment also received the highest mean class score. Such open-ended assessments, when properly guided, provide students with clarity about the requirements and expectations (Wiggins, 1990). Second, such tasks also enable students to take more active roles in the assessment process. Totoro's guided fieldwork assessment task, presented in Figure 5.1, illustrates these two features: she provided scaffolding for the task, and her students had to organize and plan their responses to the fieldwork task. Such active construction and organization of knowledge is more engaging for students than if they merely reproduce what the teacher has formerly said (Newmann, et al., 1995). A spontaneously penned reflection by Totoro's students illustrates how assessments that are thoughtfully and purposefully designed are able to engage students. It illustrates effectively how students can be so deeply enthused and energized by the assessment task that they ask for more of such activities.

> Through this trip to XYZ Hydroponics Farm, both of us felt that not only had we learnt more about hydroponics farms, but also matured further in our interviewing skills and on how to ask relevant questions and select the correct answers. Initially, both of us thought that the problems or challenges farmers face was only the shortage of money for maintenance and many more. However, after the fieldtrip, we realized that not only do farms face financial problems at times, but also the shortage of workers. One thing both of us were shocked at was the amount of charity work XYZ farm does to reach out to the society. Both of us feel that XYZ Hydroponics Farm has gone beyond the duty of an

> average farm. It has a heart to serve and give back. We both feel that this has inspired us a lot and we will strive to do likewise when we grow up and enter the work force. This fieldtrip was really enriching, fun and of course educational. It was a really great experience for us. We do look forward to more of such fieldtrips and outings.
>
> (SCH006HA6 and peer).

This unsolicited reflection powerfully captures the essence and spirit of lifelong learning that undergirds the TSLN vision of developing citizens with the passion and willingness to continue learning, not just for professional advancement but for personal enrichment.

While it is important for teachers to use assessments that challenge students intellectually, this does not mean that all assessments need to reach or attain the highest levels in the AIW scale. Assessments need to emphasize different cognitive demands. Knowledge and concepts form the basis of learning, which is the reason why Newmann and Associates (1996) included *Disciplined Inquiry* among the AIW criteria. While they do not dismiss the value of acquiring knowledge and facts, they advocate that knowledge not be assessed in a discrete and disconnected manner. This is because "usable knowledge" is more than a collection of unrelated and disjointed facts (Bransford, et al., 2000, p. 9). What the data suggest are that if teachers want to prepare students for the test of life, they need to provide opportunities more frequently for students to demonstrate their mastery of a larger range of cognitive skills. Teachers can achieve this by paying more attention to the cognitive domains (or assessment objectives as indicated in the geography syllabus) set up in their assessments. To this end, the challenge in the assessments does not lie in the format (i.e., constructive-response or objective) but in the nature of the prompts. Maryanne's post fieldwork assessment (Figure 5.2), while adhering to the test blue print to some fidelity, merely confined her students to showing their ability to recall factual information about the rainfall amounts in a rainforest. In comparison, Totoro challenged her students to analyze issues regarding the high technology farm's organization and structure (Figure 5.1).

Based on the assessments submitted for this study, while the eight teachers assessed some higher-order skills, the larger proportion of their assessments only focused on knowledge and facts. With the exception of the three project tasks which used prompts requiring students to apply and use higher-order cognitive skills, the other assessments were less aligned with the assessment objectives stipulated in the syllabus. In short, the nature and quality of the assessments submitted diverged from the TSLN intent. More importantly, these assessments limit the way students engage intellectually and culturally with the curriculum (Koh & Luke, 2009), since the prompts only require students to interact superficially within the nature of the discipline. Therefore, these eight teachers should more frequently include prompts that provide opportunities for students to apply knowledge purposefully and meaningfully (Newmann & Associates, 1996; Wiggins, 1990). To this end, one must ask how teachers guide and steer their students to enhance their learning. More specifically, as in the overarching research question, based on the student responses on these assessments, one can also ask how these teachers make formative use of assessment data to enhance student learning. These questions are discussed in Chapter 6.

CHAPTER 6:

TEACHER ASSESSMENT AT THE CLASSROOM LEVEL (2012): QUALITATIVE ANALYSIS

Introduction

After examining the types of assessments the eight teachers participating in this study submitted, the goal of this chapter is to understand how and why these teachers construct, use and review their classroom assessments in order to to answer the overarching research question:

> **Under an educational policy that emphasizes the preparation of students for "the test of life" instead of a "life of tests" (MOE [Bluesky], 2005), how do Singapore geography teachers elicit and enhance student learning through the ways they use classroom assessment?**

The three sub-questions addressed in this chapter are:

- What does "assessment" mean to Singapore geography teachers?
- After implementing their classroom assessments, how do Singapore geography teachers make formative use of assessment data?
- What factors influence the nature and quality of classroom assessments designed by Singapore geography teachers in response to the *Thinking Schools, Learning Nation* vision?

The qualitative interview data presented in this chapter provide insight into three different aspects of the teachers' assessment practices. First, the teachers' views of "assessment" provide a means by which to understand the assessment practices they enacted. Next, the teachers' formative use of assessment information is a way to examine how they supported and enhanced student learning after marking and analyzing students' responses. The final section of this chapter presents the factors that influenced the classroom assessments that the teachers submitted.

In line with the explanatory mixed methods design of this dissertation, the qualitative data are used to explicate or elaborate on the quantitative results presented in Chapter 5. The combination of multiple data sources, and analyses and interpretation approaches enables the triangulation of research findings (Greene, et al., 1989) in order to examine the different ways in which Singapore geography teachers elicit and enhance student learning through the ways they

conduct classroom assessment. To this end, this chapter integrates these teachers' qualitative perceptions into the quantitative classroom level data and makes "meta-inferences" to bring about "increased *Verstehen* (or understanding) (Onwuegbuzie & Combs, 2010, p. 398). In this chapter, "data comparison" (Onwuegbuzie & Combs, 2010), one of the "cross-over strategies" for mixed methods is used to compare the qualitative interview data reported in this chapter with the quantitative AIW findings presented in Chapter 5. Chapter 7 will apply another cross-over strategy, "warranted assertion analysis" (Onwuegbuzie & Combs, 2010) which involves the use of all data sources to arrive at meta-inferences in response to the overarching research question.

Background and method

The qualitative interview data were grouped using theoretical codes which involved categorizing the data into a framework based on prior theory (Maxwell, 2013). Two different lenses—theory and policy—were used to categorize the teachers' comments.

Theoretical lens

Since TSLN and TLLM resonate closely with constructivist learning theory, the analysis of the teachers' interview data adopts this theoretical lens. Shepard's (2000) vision of assessment practices in an "emergent constructivist paradigm" is used to code the teachers' comments (Figure 6.1). The key features of classroom assessment in Shepard's framework (2000, p. 8, shown in Figure 2.1) below closely mirror the TLLM spirit and intent.

Figure 6.1
Features of constructivist classroom assessment (adapted from Shepard, 2000)

1. Presents challenging tasks to elicit higher-order thinking
2. Addresses learning processes as well as learning outcomes
3. Is an on-going process that is integrated with instruction
4. Is used formatively to support student learning
5. Is used to evaluate teaching as well as student learning
6. Makes expectations visible to students
7. Involves students in evaluating their own work actively

Constructivist classroom assessment in Shepard's (2000) framework draws on cognitive, constructivist, and sociocultural theories. It aims to find out "*what* the learner knows, understands or can do" (Torrance & Pryor, 2001, p. 617). Comparatively, assessment from the traditional, behaviorist paradigm seeks to elicit "*if* the learner knows, understand or can do a predetermined thing" (emphasis as in the original, Torrance & Pryor, 2001, p. 617). To this end, one key feature of constructivist assessment is the **teacher's assessment and learning goals** (Figure 6.1, Points 1 and 2). Furthermore, the use of challenging tasks to assess higher-order thinking resonates with Newmann and Associates' (1996) authentic intellectual work in calling for classroom assessment tasks to elicit thinking and learning processes (Shepard, 2001).

A second feature of constructivist assessment is **formative assessment** (Figure 6.1, Points 3, 4 and 5), which is assessment conducted during instruction to inform teaching and support learning (Shepard, 2006). Formative assessment includes the use of informal methods such as observing or open questioning of students, or the formative use of formal assessments like tests (Shepard, 2006). The constructivist paradigm envisages the frequent use of formative assessment to provide feedback to students to move them from their current level of performance or achievement towards the intended goal (Hattie & Timperley, 2007; Ramaprasad, 1983; Sadler, 1989, 1998). Frequent assessment enables teachers to check constantly on their students' understanding (Pellegrino, et al., 2001). As mentioned in Chapter 2, formative assessment in this

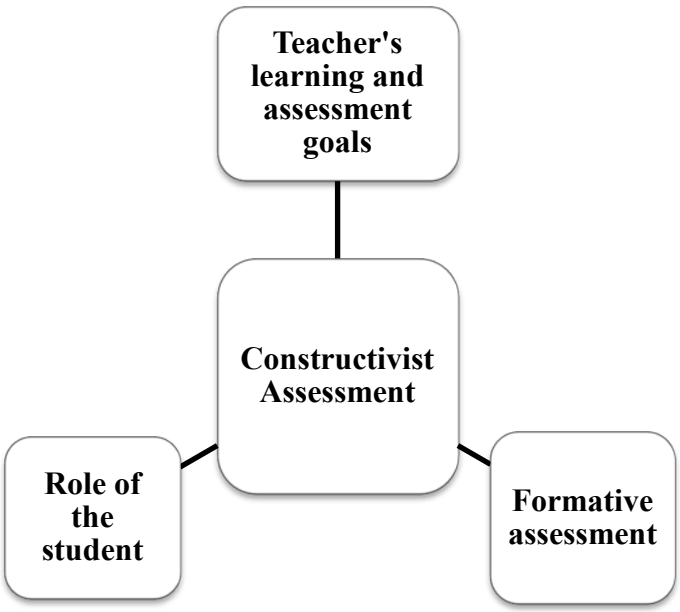
study refers to "all those activities undertaken by teachers, and/or their pupils, which provide information to be used as feedback to modify the teaching and learning activities in which they are engaged" (Black & Wiliam, 1998a, pp. 6-7). This means that formative assessment goes beyond written tasks and assessments to include classroom activities. Furthermore, this definition also suggests that summative assessments can also be deemed formative if teachers used student achievement data from these assessments formatively to inform and improve their future teaching plans. Hence, the interview protocol included questions that asked the teachers to interpret the work their students had submitted, and to discuss the follow up teaching decisions that they would take (Appendix 3).

The third feature of constructivist assessment is the **role of the student** during assessment (Figure 6.1, Points 6 and 7). Black and Wiliam (2009, p. 9) refer to formative assessment as classroom practices in which "evidence about pupil achievement is elicited, interpreted, and used by teachers, learners or their peers, to make decisions about the next steps in instruction." In constructivist assessment, students participate actively because they have to be aware of the goals they need to achieve in relation to where they currently stand (Black & Wiliam, 2012b; Ramaprasad, 1983; Sadler, 1989). This means they incorporate the information given to them, and use this to plug gaps in their learning. In so doing, students construct and build on knowledge, and change from being passive to active learners. Therefore, peer and self-assessment are important aspects of constructivist assessment as they provide opportunities for students to be actively involved in learning and assessment. Peer assessment is particularly valuable because by examining their peers' work, students can better understand the standard of their own work (Black & Wiliam, 2012a). As students were not interviewed for this study, and

no classroom observations were conducted, the role of the student in classroom assessment is derived from the teachers' accounts of how the assessments were enacted in the classroom.

In summary, the theoretical coding and presentation of the teachers' classroom assessment practices were based on three features of constructivist assessment derived from Shepard (2000), namely: (1) assessment and learning goals, (2) formative assessment, and (3) the role of the student, as shown in Figure 6.2.

Figure 6.2
Adapted features of constructivist assessment



Policy lens

The second lens used to categorize the teachers' comments is the policy lens. The teachers' interview comments were categorized based on the extent to which their assessment practices resonated with the TLLM tenets (see Table 2.1). Specifically, TLLM encourages teachers to do "*more* formative and qualitative assessing," "*more* guiding, facilitating, and modeling," "*more* process," "*more* for the test of life," "*more* for understanding," and "*more* searching questions" (MOE [Bluesky], 2005). TLLM also exhorts teachers to do "*less* telling,"

"*less* summative and quantitative testing," "*less* product," "less for a life of tests," "less dispensing of information only," "less textbook answers," and "less set formulae, standard answers" (MOE [Bluesky], 2005). The policy's use of "more" and "less" signals a re-calibration of and re-balancing in emphases rather than a swing from one type of teaching to another.

Organization

One emerging pattern from the analyses of the interview data was that the teachers fell into different groups in terms of how they developed and used assessments to elicit and enhance student learning. Some teachers' classroom assessment practices were *more*, some *moderately* and some *less aligned* to TSLN and TLLM. As I have argued that these two policies are underpinned by constructivist theories, the three categories can also be interpreted as the teachers' assessment practices being more, moderately, and less reflective of constructivist assessment.

More aligned teachers. The first category is comprised of teachers whose practices were *more aligned to TSLN*. As shown in Table 5.8, the assessments that Harry and Totoro submitted received high AIW scores because the skills tested and prompts used mirrored the TSLN intent. Their learning and assessment goals, and the way they enacted formative assessment closely mirrored the TLLM tenets and constructivist learning theories.

The two *more aligned* teachers, Harry and Totoro, have taught for over ten years. Harry worked in a public school while Totoro taught in an independent school. At the time of the research, Harry was teaching Secondary 2 students while Totoro was teaching Secondary 1 classes. Both teachers held leadership positions in their schools.

Moderately aligned teachers. The four *moderately aligned* teachers – Amanda, Jiajia, Margaret, and Miki – taught in public (Jiajia and Amanda) and government-aided (Miki and Margaret) schools. Both Amanda and Miki had taught for over ten years while Jiajia had about

five years of teaching experience. Margaret was the youngest teacher and had been teaching for about two years. At the time of the research, Amanda, Jiajia, and Margaret were teaching Secondary 2 students while Miki was teaching Secondary 1. While Miki and Margaret taught in single-sex schools, Amanda and Jiajia worked in co-educational schools. Amanda, Jiajia, and Miki had AIW scores in the middle of the AIW scale. While Margaret's scores were towards the lower end of the AIW scale, her discussions about her assessment practices resonated more with the views reflected by the *moderately aligned* than with the *less aligned* teachers.

At times, the *moderately aligned* teachers expressed comments about assessment, higher-order skills, and formative assessment that were aligned to TSLN's goals. At other times, they spoke of and adopted practices that diverged from the TSLN and TLLM policy intent. While these teachers valued higher-order skills like analysis and effective written communication, their purposes for meting out such assessments were to prepare students for the upper secondary high-stakes examinations. Thus, while they said that they emphasized higher-order skills, their goals were related to passing examinations rather than developing skills to prepare students for life.

Less aligned teachers. Teachers in the third category were *less aligned to TSLN* because their assessment approaches and comments tended to emphasize surface learning (Marton, Dall'Alba, & Tse, 1996), which involves the rote recall of facts without making connections among the pieces of knowledge. These two teachers focused on ensuring that students produced the correct answers rather than guiding them to understand the purpose of the learning.

The *less aligned* teachers, Maryanne and James, taught in public schools. An experienced teacher, Maryanne had been teaching geography for 19 years. James, on the other hand, was a novice teacher with about three years of teaching experience. Both Maryanne and

James taught students with similar profiles to those taught by Harry (*more aligned* teacher) and Amanda (*moderately aligned* teacher).[20]

The *less aligned* teachers' comments and assessment approaches were aligned with those practices that TLLM urged teachers to use less of. Both teachers received AIW scores that were below the scale midpoint, meaning that their assessment tasks did not address the higher-order skills aligned with TSLN's goals. Their assessment practices focused on test strategies and examination preparation, and thus, deviated from the policy's intent.

Table 6.1 presents the teachers' grouping in the categories in relation to their AIW scores. Generally, the teachers' AIW scores matched the extent to which their views and practices were aligned with TSLN:

- The *more aligned* teachers tended to have high AIW scores, suggesting that their assessments focused on higher-order cognitive demands. Their scores and interview comments also reflect close alignment with the goal of preparing students for life after school.
- The *moderately aligned* teachers' interview comments suggest that while they focused on both the need to prepare students for the *test of life* and for a *life of tests*, their predominant goal was to obtain good test and examination scores, especially for the upper secondary high-stakes examinations.
- The *less aligned* teachers received low AIW scores indicating that their assessments tended to focus on the reproduction of knowledge and on the preparation of students for examinations as they spoke frequently in the interviews about practice tests and model answers.

[20] This is based on the Primary School Learning Examination scores of students entering the schools.

Table 6.1
Teacher groupings and AIW ranking

Teacher	Alignment to TSLN	Mean AIW scores (*SD*)[a]	AIW Rank
Harry	High	17.3 (1.2)	1
Totoro	High	17.0 (2.0)	2
Miki	Moderate	15.0 (1.0)	3
Jiajia	Moderate	14.7 (1.5)	4
Amanda	Moderate	13.0 (1.7)	5
Margaret	Moderate	12.3 (3.8)	7
James	Low	13.0 (1.7)	5
Maryanne	Low	9.2 (0.8)	8

[a]Same mean AIW scores as in Table 5.7.

As mentioned in Chapter 5, at the lower secondary level—the level at which these eight teachers teach—there are no high-stakes assessments. All assessments, including the end-of-the-year summative assessment are school-based, and schools have the autonomy to decide on the frequency and nature of the assessments. Since the launch of TSLN in 1997, the call has been for teachers to assess higher-order skills. As a result, the geography syllabus uses Assessment Objectives to specify the balance in weighting between the assessment of knowledge and higher-order skills.[21] The syllabus also calls for teachers to use a range of assessment types, including portfolios and fieldwork assessments.

Therefore, one must ask how these teachers have responded to the TLSLN and TLLM reforms, what their class assessments are like, and why they use such assessments. The eight teachers have been categorized into three groups based on the frequency and extent to which their interview responses resonated with or diverged with the TLLM and TSLN intents. Grouping the teachers enabled comparisons to be made. As I pointed out earlier, the TSLN and TLLM policies resonate with constructivist theory. Thus, the teachers' comments are interpreted,

[21] As mentioned in Chapter 5, the syllabus uses Assessment Objectives (AOs) to inform teachers to assess higher-order skills. The stipulation is for Knowledge (AO1) to be assessed together with Critical Understanding and Constructing Explanation (AO2). Knowledge (AO1) should also be assessed with Interpreting and Evaluating Geographical Data (AO3). This means that teachers should not assess discrete knowledge and facts, but test knowledge together with higher-order skills.

analyzed, and presented using the three themes of constructivist assessment depicted in Figure 6.2. To this end, this chapter provides details and explanations of the classroom assessments that the eight teachers used to elicit and enhance the learning of their Secondary 2 geography students.

More aligned teachers and classroom assessment

The assessment practices of the two *more aligned* teachers resonated most with the TLLM tenets. To a large extent, both teachers enacted classroom assessments that reflected the TSLN and TLLM visions to the spirit and to the letter. As compared to the diversity of practices among the four *moderately aligned* teachers, the practices of the two *more aligned* teachers converged. Both teachers had similar views in the three areas of constructivist assessment, namely assessment and learning goals, formative assessment, and the role of the student.

Assessment and learning goals

As shown in Figure 6.2, one aspect of constructivist assessment is the teachers' assessment and learning goals. The goals of the *more aligned* teachers were aimed at the long term. Harry and Totoro frequently talked about preparing their students for life after graduation. They wanted their students to be able to apply and construct knowledge; to demonstrate the ability to take multiple perspectives; to engage in processes of analysis, evaluation, and interpretation of information; and to apply knowledge to new or real world contexts. In short, their assessments closely mirrored the AIW criteria. They spoke frequently about the complexities of the world, and emphasized that their goal was to help their students live in and contribute to society. As a result, these teachers' assessment and learning goals went beyond ensuring that their students accumulate knowledge and pass examinations to helping their students to develop as active citizens and to be ready for life outside of school.

For the *more aligned* teachers, the "end goal" of learning and assessment "has to be the kind of the skills that they need to learn" (Totoro). Since life in a knowledge-based society will be "fluid" and "answers are not very clear," for Totoro, the purpose of 21st century education is to help learners "be comfortable with ambiguities." In her view, students needed to know how to consolidate and interpret the many sources of information, and then justify or explain the phenomena based on their evaluation of the evidence. Her goal was for her students to "be comfortable with unfamiliar environments." This was evident in the design of the fieldwork assignment and in the way she required them to integrate various sources into their work and in the way she guided them to evaluate the veracity of information. She pointed out that

> Ambiguities can come … in class … The textbook says that the government will assist the high tech farms here to be sustainable and things like that. But when they visited XYZ farm, it was not so, especially for the first 2, 3 years when the farm was first started. According to the farm guides, that there was very little help from the government. But then they say this is textbook knowledge. But the real world knowledge tells me that. But I'll tell them, 'when you read this, because now you are given a context specific problem and a challenge. Does it mean that the textbook is wrong? What kind of timeframe are we talking about? Now you think about. If every high-tech farm the government has to help and ensure them, then where are the business competencies that you are talking about? But eventually, did the government not help the high tech farm?' This is something that gets them to think. Because the textbook is meant for Sec 1 students, [it's] simply condensed. So the girls think everything's very simplistic. But when they do fieldwork, things become more complex.

Totoro's comments reflect many TLLM tenets. First, she did not want to tell her students the right answer but rather guided their thinking—a practice resonating with the TLLM tenet of teachers doing more guiding, facilitating and modeling. Second, she was prepared not to have standard answers so that the assessment could be authentic to the task. Third, she challenged her students to question the textbook's presentation of the situation in relation to their experiences and observations at the farm. Her approach to guiding her students to learn also closely reflects AIW criteria such as *Construction of Knowledge* and *Value Beyond School*.

While the *more aligned* teachers spoke of assessment goals that focused on developing higher-order thinking and preparing students for life outside of school, this is not to say that they were not burdened with the accountability aspects of assessment. They were cognizant of the need to be accountable to stakeholders. Hence, they needed to "help those who don't score the perfect score" (Harry) to improve their marks. In Singapore's "competitive society" parents want to know "how my girl is doing" (Totoro). As a result, these teachers were also mindful if they "see too many reds, too many single digits" (Harry). At the end of the day, "we have to be very practical" (Totoro).

However, what differentiates these two teachers from the *moderately* and *less aligned* teachers is that they used assessment scores purposely and pedagogically. Test scores directed Harry to question himself, "Am I a good teacher?" Totoro indicated that "summative [tests] can be formative in nature" because these are "ongoing processes … summative can become formative, and formative can become summative." The teachers' comments resonate with Shepard's (2000) vision of constructivist assessments in which assessment needs to evaluate and provide information in order to improve teaching in addition to supporting student learning. In contrast to colleagues in the other two categories, these two teachers did not mention retests or practice tests in any of their three interview sessions. This is because the *more aligned* teachers viewed assessments as once-off events, or as Totoro pointed out, "just a snapshot." Both teachers believed strongly that "all students can learn" (Harry). As a result, for Totoro, a low score would not mean that

> she cannot be a very bright geog[raphy] student, say, four years down the road. I will always tell them that at this point, I measure you as such now. Your potential has not been realized, or you may not be ready. I just make sure that in this journey, you just move on and then cognitively you're slowly building up. It's a journey, a process.

Therefore, the *more aligned* teachers did not view students' abilities as fixed, but they embraced the belief that all students can develop after being guided in learning. Given their learning goals, it is not surprising that the *more aligned* teachers employed assessments that sought to elicit higher-order skills from their students. Compared to their *moderately* and *less aligned* colleagues, the *more aligned* teachers were less preoccupied with the acquisition of knowledge. They believed that assessment should not focus on content because "today, in a world of IT, if I don't know certain facts, I just Google" (Totoro). Therefore, Totoro's assessments were to "test thinking, rather than how well you memorize."

The emphasis that these two teachers placed on higher-order skills in their assessments was evident in their AIW scores. As shown in Table 6.1, both Harry and Totoro had the highest AIW scores. As Totoro explained, the assessment goal was to ascertain if students were able to adopt a "geographical lens when looking at issues or problems, whether at the school level, whether at the national level, or better still, look at some of the global issues." Similarly, Harry's goal was for students to understand that "geography is every day." As a result, the true test as to whether students had understood the topics discussed in class or whether they had grasped the big idea was when they read the newspapers and then come up to say, "Sir, we've done this in class." These views of teaching and assessing geography resonate with the AIW criteria—*Construction of Knowledge*, *Disciplined Inquiry* and *Value Beyond School*—since these teachers wanted their students to be able to interpret and explain patterns and phenomena occurring in the world around them based on the knowledge, and information that they are given or have been taught.

A distinguishing aspect of these two teachers' assessment is the attention paid to the *consideration of alternatives* standard (from the *Construction of Knowledge* criterion). Of the

eight teachers, only Harry and Totoro addressed aspects related to this AIW standard in their assessments and in their interview comments. Harry did not want his students to rely solely on the textbook as the source of truth since there are many perspectives. When planning his assessments, he checked, "Did you follow the textbook and everything else? … [he did] not want the test to be entirely based on textbook material." This explains why in the two research assessments, Harry required his students to collect a variety of data sources and to provide several viewpoints. Under "What to include" in the task, he reminded his students to consider "Specific examples and names," "What has happened? What will happen?" "Government's views," and "Maps, diagrams, tables, figures." Combining data sources enabled students to analyze issues from different perspectives. To guide his students, Harry never failed to remind them that they must "always have both sides" because, "every issue always has several sides." After much practice, Harry's Secondary 2 students were then able to provide responses like, "I agree, but …" To Harry, "*but* is the key word" and he would guide his students to "counter argue." In this sense, he wanted them to develop the ability to appreciate different perspectives.

Similarly, Totoro's assessment goals were for her students to be able to "make their own conclusions, their own decisions." This is why in the fieldwork assessment Totoro challenged her students to produce responses that went beyond describing the farm processes. She wanted them to "discuss the opportunities and challenges faced by the farm" and to identify "factors that that influenced the decision to adopt hydroponic technology for growing vegetables." Because information sources are diverse, Totoro was cognizant that "it is not possible to have one standard response," especially if the questions are on "analysis" and "evaluation." As such, she had a flexible marking scheme for the fieldwork assessment because she was aware that her students gathered a wide range of data, depending on the types of questions they posed at the

farm, and the types of responses provided by the farm employees. These responses provided an alternative to the information furnished in the textbook, and the flexibility Totoro exercised in the marking made the assessment authentic to the task. She explained that

> A typical geography answer may be they just talk about leafy vegetables such as cai xin [local vegetable], or herbal spices here such as basil. But in fieldwork, the girls go beyond and tell you things like the distribution channels to Cold Storage and Shop 'n Save. It won't come out in the geography textbook. This is something from the farm itself, in terms of the amount they produce. They tell you about when they have output here, they would package them as 250g and then selling it at $1.50. You can't get this in your pen-and-paper.

These two teachers used prompts and task formats which resonated with their views of assessment. As reflected in their AIW scores (see Table 5.8), these two teachers did not focus on discrete and disjointed bits of knowledge, but included tasks that required students to relate theory to real world contexts. Totoro's fieldwork assessment (Figure 5.1) asked students to assess the potential of high-technology farming in contributing to Singapore's food supply. In another assessment for her Secondary 1 students, Totoro presented her students with charts, tables, pictures, and text, and required her students to interpret "different statistics" based on a variety of world contexts, from water use in India and the world to the sources of water supply in Singapore, and then to explain the patterns using disciplinary content or to make projections of future trends. The distinguishing characteristic is that these assessments went beyond the testing of learned knowledge to requiring students to "organize and present information in a coherent manner," (CPDD, 2005, p. 1) and to explain these phenomena in relation to geographical concepts. These skills strongly resonate with the AIW criteria, and by extension, with the teachers' goals of preparing students for life outside of school.

Based on their learning and assessment goals, the *more aligned* teachers used a variety of assessment formats and types to elicit student learning. Of the six assessments these two teachers

submitted, just one assessment—Harry's common test—used the format of the test blueprint similar to that suggested in the syllabus. The other five assessments submitted by these two teachers were research projects, and data interpretation exercises. Using a wide variety of assessment formats enabled the teachers to elicit different types of learning. Unlike the *moderately* and *less aligned* teachers, Harry and Totoro assessed both the process and the outcome. In addition to presenting his students with research projects that required them to collect, organize, and interpret information and data, Harry also tasked them to reflect on the projects they had completed. These reflections focused on the learning process during the independent research study, as well as what students had gleaned from the research. Figure 6.3 shows two excerpts from Harry's students.

Figure 6.3
Student reflections

Student A	Student B
In my perspective, I think that there are many other possible challenges that may occur close at hand, and we should not get too complacent despite all the facilities given to us by the government as the future is unpredictable. We should not be too dependent on others and be well-prepared if anything were to happen. Even though Singapore is a small country, we should do our part as Singaporeans to maintain law and order and also make it a best place to live, work and play.	I have learnt that if these Flash Floods do not stop, people would get sick of it and might migrate to other countries. The government should take immediate actions such as building more drains etc. The Flash Floods might affect us, [our school] as was one of the places that had been affected. Quite a lot of [students from our school] travel through [this road] to go to school so if a flash flood had occurred, the student would be affected and be late for school.

The extracts show that Harry's students reflected on different experiences based on their chosen topic. Student A's reflection focused on the need to be prepared for calamities. Student B focused on the daily commute to school and how the lives of fellow students in the school would be affected by flash floods. In getting his students to reflect on the task and the issue, Harry was ensuring that his students pondered deeply over the task they undertook, and this

increased their understanding of the purpose of the assignment. The use of reflections mirrored Harry's assessment goals for his students which were to relate classroom content to their daily lives and experiences. Like the spontaneous reflection Totoro's student submitted, students could be engaged in tasks that they found meaningful and purposeful.

Formative assessment

The second aspect of constructivist assessment (Figure 6.2) is formative assessment, which includes practices integral to teaching and learning, as well as teachers' formative use of assessment information (Shepard, 2006). Formative assessment is a critical leverage point in classroom practice because the interaction between teachers and students during formative assessment is at the "heart of pedagogy" (Black & Wiliam, 1998a, p. 16). The formative use of assessment data involves teachers interpreting evidence from student work and from interactions with students, and then making decisions as to the next steps in instruction (Black & Wiliam, 2009). This use of data is formative because teachers are able to make better curricular and instructional decisions than they are able to do in the absence of the assessment information (Black & Wiliam, 2009). Drawing on Shepard (2006) and Sadler (1989), the analyses of the eight teachers' formative assessment practices focus on how they made formative use of assessment information as well as the formative assessment strategies they employed.

Formative use of assessment information. Teachers' effective use of assessment information is critical to improving students' learning. Formative assessment involves teachers examining and making decisions about the quality of student work which is then used to shape and enhance the student's learning (Sadler, 1989). According to Sadler (1989), formative assessment has three important features. First, students and teachers need to have a clear idea of learning goals or what Sadler (1989, p. 121) terms the "reference level." Second, there must be

information, evidence or data that pinpoints the student's present level of achievement. Third, there must be action to close the gaps between achievement and goals in order to move students towards the intended outcomes.

Based on Sadler's model, after eliciting the teachers' goals, the next step is to examine how they interpreted students' learning and performance after each assessment. This enables us to understand how teachers ascertained their students' current level or state of achievement before deciding how to close gaps in students' learning. After a thorough analysis of her students' work, Totoro would identify "What is the typical problem? Then I will color-tag it. So I have four questions—four different color tags. So I roughly know that for this one, this is the typical problem!"

For the *more aligned* teachers, formative use of assessment information involved looking out for what students had shown they are able to do over time, rather than merely base their judgment on the scores of one assessment. As mentioned above, Totoro was mindful that each assessment was a snap shot of the learner's performance because her Secondary 1 students "can grow and develop." For her, when students "do well in the assignment, they have done the learning based on the learning goals that you have put in your assessment questions." In a similar way, Harry focused on students' development over time. He encouraged his students to create a portfolio of their work completed over a the course of school year because "you want to show your parents this is what I've done in class. Your file represents you. Your file represents the teacher."

Another common practice was to identify students' strengths in the piece of work, and then to make decisions in order to further stretch students. Totoro noticed that since her "students can elaborate and explain. It's very good! Now we can go beyond what has been said."

Thus, formative assessment for Totoro means building on what students have shown they are already able to do. This approach is not mentioned by the *moderately* and *less aligned* teachers. When examining her students' responses to the data response questions, Totoro was delighted that her students were able to "see the link between the annual water usage and the population. And when they explain it, you can see that they've understood." Based on her analysis of her students' work, Totoro was certain that her students "can be accelerated" and decided to create more challenging tasks for them in order to "add value to their learning." Subsequently, she decided that "since the general cohort seems to have grasped it, instead of having just one variable in the question, they may now see two variables." This would enable her to challenge the higher-ability students so that they "will not find it so boring" as well as "expose the middle ability students."

Harry adopted a similar approach. Based on his students work, he would question himself, "Am I stretching them? How much more can I stretch you?" Therefore, he did not believe in giving his students "too much notes" but would prefer that his students "sit down and write their own notes because you become clearer in your learning." Furthermore, when students made their own notes, their responses in the examination were "less rigid." At the end of the day, he did not want his students to be "fervent followers of the textbook." This approach to teaching and assessment was very different from that of the *moderately* and *less aligned* teachers.

The *more aligned* teachers more frequently focused on students' performance in skills rather than content. Harry spoke about how his students were able to organize their writing using terms like "the purpose of this essay" and "I'm going to elaborate." Likewise, when discussing her students' performance in the third assessment, Totoro discussed her students' weaknesses in not being able to "see links across what they learn." She observed that her

students were looking at "each idea in isolation." Another weakness was that students needed to understand the difference between "elaboration and explanation" and the link between "the causes and the effects." In her opinion, such skills were necessary for students to write well and communicate their ideas. Otherwise, they would not be able to present "a more balanced view."

When they saw deficiencies or gaps in student work, the teachers would first question themselves and reflect on their teaching. They would use the information to review how and what they teach. Harry would analyze his students' responses and question himself, "How have you taught the class?" He believed that teachers should not merely say that "students do not know or students are not very intelligent." Instead, a "teacher has to relook and say how much have you taught correctly. Otherwise, we have to relook at certain things."

Formative assessment strategies. Based on Sadler's model, formative assessment serves to close gaps in students' learning by moving them from their current level of performance towards the intended reference levels. To do so, they need to know how to move towards the intended learning goals or outcomes. The *more aligned* teachers used a variety of formative assessment strategies to inform teaching and learning, including questioning, observing, and having discussions with students.

When giving feedback, these teachers frequently used open questioning to probe and guide their students toward understanding. Their use of such approaches resonates with "divergent assessment" (Torrance & Pryor, 2001, p. 617) which is associated with constructivist views of learning. As the *more aligned* teachers had practices similar to the TLLM tenets, they did not immediately tell students the right answers when reviewing an assessment. Harry would not advocate "flashing" out the right answers for students to copy. His preference was to make his students work out the answers themselves. Based on his fifteen years of teaching experience,

Harry said that when examination and common test papers were returned, "students are only keen on knowing their marks. They would not focus on what they copy," and hence, would not learn. As a result, he would only give "general comments" and then he would ask his students to "focus on the response, or to tell [him], how do you feel about it?" Likewise, Totoro encouraged her students to analyze and figure out the errors. One method that she used to help them understand their errors would be to "give them a sample," and "ask them to compare this with their own work." Through this method of feedback, she observed "they actually learn something" because "when they figure it out for themselves, it is more powerful than when I tell them." Giving exemplars and samples to students was one way to "try to build the blocks." As a result, she would "give them a sample of the good answers so that they can model" and compare their responses to the exemplars.

When time permits, Harry would devote a substantial amount of time to "mark and re-mark" his students' work and to give feedback. In fact, he "keeps marking." For the first assessment which the students completed over the one-week March break Harry spent a substantial amount of time with his students giving them feedback and allowing them to re-draft because he "really work[s] with [his] students … and if [he] had the time … to re-do their essays." As compared to the *moderately* and *less aligned* teachers who led their students towards one specific answer, Harry tended to highlight his students' strengths and weaknesses. One example was to inform the student that "you have looked into each challenge with consideration. Well-crafted and mature thinking and formulating your report." He would provide suggestions for improvement, such as, "You need more detailed solutions like what and how the government is helping." The nature of this feedback explains the higher AIW student mean scores for one of the two independent research assignments he submitted. However, revising and re-drafting are

only possible if students cooperate and "understand the meaning of submitting early" as this would enable him to have "time to re-look, and give feedback." As a result of feedback, the revised pieces of work were "longer responses" and students were "taking more critical stances."

Recognizing that their students had different abilities and strengths, and that each group of students was different, the more aligned teachers tried—when possible—to provide differentiated and individualized feedback and they used a variety of approaches. Harry would provide individualized feedback for each student and have them re-draft his or her independent research. Totoro, too, tried to provide customized feedback. However, because of the time constraint, she only applied this approach with the high-ability students. In addition, after returning an assessment to her class, Totoro would ask her students, "Why is this answer wrong?" She firmly believed that this type of questioning was important because "if they can tell [her] as a class … they actually learn something." Sometimes she would provide her students with anonymized copies of students' work and then she would ask her students, "How can you make it better?" as this is "more challenging for the better class." She observed that

> They very quickly say why this is a good answer. The better students, they can tell, oh, ok. And then they ask, "Why is it that I got this?" Then I ask, "What is it that is missing?" They can tell themselves—good! You've given your own feedback.

Totoro only used this feedback approach with better classes because they can "move on quite fast, accelerated." Conversely for the "not so good class," she would "just tell them what's wrong." In the example above, Totoro's used open questioning as a feedback strategy. She would not immediately provide her students with the response. Instead, she would guide her students to think through their responses.

Role of the student

The third aspect of constructivist assessment (Figure 6.2) is the role of the student. As I suggested, the TLLM tenets envisage an active student role in learning and assessment. This view resonates with constructivist theories because the dominant role of the teacher as a sage-on-the-stage has now morphed into a collaborative partnership with students during the learning process (Shepard, 2001). Constructivist assessment also envisages that students actively reflect on their work based on their understanding of the learning goals (Shepard, 2001).

The *more aligned* teachers tried, when possible, to enable their students to participate more actively in class. Using open questioning techniques was one way to provide feedback to nudge students to think and to reason for themselves. This was Harry's approach when he gave feedback to a questionnaire his students were developing to study 'Pollution.'

> We talked about interview skills. They said, if all my questions are "no," very negative. Then I asked them, "What's wrong?" And then they said, "The way I construct the questions." I said, "You want to elicit responses. And how do we elicit responses?"

Similarly, Totoro used broad questioning techniques to direct her students towards self-reflection. For example, as they rode the bus back to school after the farm visit, Totoro capitalized on the recency of the experience to ask her students to reflect on the interviews they had conducted. Rather than pointing out that some of them had asked "bad questions," Totoro invited her students to discuss the experience of interviewing, and to identify what they thought could be improved on. Her students were so spontaneous in identifying "bad questions" and learning points that Totoro and her colleagues were amazed that "they can remember" large parts of the discussion with the guides and the farm employees. Such a learning experience was made possible because as students were interviewing the farm staff, Totoro did not intervene or

interject. Rather, she took a step back and allowed her students to take charge by inviting them to ponder and reflect on the experience. In short, she drew the learning out of them.

In constructivist assessment, students are envisioned to play more active roles in the classroom (Black, et al., 2003b; James, 2006; Pellegrino, et al., 2001). Through strategies like peer and self-assessment, students were able to develop an idea of the expectations and goals and take action to achieve these goals. However, even among the *more aligned* teachers, there were limited examples of students playing more active and participatory roles during classroom assessment. Even Totoro only made this opportunity available to the higher-ability classes. Neither teacher provided opportunities for peer or self-assessment.

Summary

The *more aligned* teachers' comments on their classroom assessment approaches mainly resonate with the TLLM tenets and TSLN's emphasis on teaching and assessing higher-order thinking skills. They provided avenues for students to figure out answers rather than be reliant on formulaic, standard answers. Over the five-month study period, while the *more aligned* teachers used pen-and-paper assessments like their *moderately* and *less aligned* colleagues, the assessment formats they used varied more widely over the course of the school year. Using a wide variety of assessments meant that their students could demonstrate what they were able to do in different ways rather than to merely respond to prompts appearing in the task. As indicated by their AIW scores, this pattern of assessment use suggests that the *more aligned* teachers' used assessments that expected their students to demonstrate higher-order skills.

The *more aligned* teachers adopted formative assessment practices that reflected the TSLN intent. They used formative assessment to inform their teaching and adopted a variety of strategies when providing feedback. Although they typically provided whole class feedback,

they tried to find time to provide group or individualized feedback that was tailored to students' needs. An important aspect of their practice was that they did not want their students to be overly reliant on them to dispense the correct answers or solutions. Hence, they used open questioning to direct their students to think and reason independently. In spite of this, the role of the student—the third characteristic of constructivist assessment—was not a feature common in the *more aligned* teachers' assessment practices. In fact, the teachers still played the dominant role in classroom assessment practices, and there was little evidence of student voice or student agency in the assessment process. There was also no evidence of students engaging in peer and self-assessment.

Moderately aligned **teachers and classroom assessment**

The *moderately aligned* teachers employed a range of classroom assessment practices that aim to prepare students for life after school as well as for examinations. Compared to the *more aligned* teachers whose views and assessment practices were homogeneous within the category, there was a continuum of practices among the *moderately aligned* teachers, with Margaret and Jiajia adopting assessment practices that were closer to the *more aligned* end while Amanda and Miki employed practices that resonated more strongly with the *less aligned* teachers.

Assessment and learning goals.

The *moderately aligned* teachers—Amanda, Jiajia, Miki, and Margaret—had assessment and learning goals that focused on subjecting students to lots of examination preparation while also developing them for the *test of life*. Of the two aspects, the teachers emphasized the former.

Like the *more aligned* teachers, the *moderately aligned* teachers adopted assessment and learning goals that went beyond the acquisition of discrete facts and knowledge to embrace the philosophical aims of geographical education. Geography assessment had to examine whether

students "understand how the world works" (Margaret). Amanda embraced "sustainability" because geography was about "how to survive in this world, given the context, given the resource that I have, given the country that I live in—the particular place." Jiajia wanted her students to "be responsible citizens of the future." Like the *more aligned* teachers, these four teachers would, from time to time, incorporate questions that required students to discuss or write about the goals of geographical education. In one assessment, Jiajia included a prompt for her students to discuss, "How do you think people can conserve and protect the environment?" Margaret asked her students to "state some of the ways you can save water at home." Similarly, Amanda required her students to "Explain why human activities have a much more powerful impact on Earth now than 50 years ago, in terms of (i) human population, and (ii) transport and communications and industrialization." Teachers assigned such prompts because "geography is earth sciences, it is studying the earth—and this is where we live" (Miki). As a result, it was "important to know how we impact the earth with what we do" (Miki).

The use of these prompts suggest that in their assessments, the *moderately aligned* teachers expected their students to apply learned knowledge to real world contexts, similar to the expectations reflected in the AIW criterion, *Value Beyond School*. However, these goals did not play out as frequently and as extensively in their assessments as in those of the *more aligned* teachers. As shown in Table 6.1, the AIW scores of the *moderately aligned* teachers were below the scale midpoint. This means that while the *moderately aligned* teachers addressed higher-order skills, they did so to a lesser extent and to a lesser frequency than did the *more aligned* teachers. The three assessment prompts presented above were the only examples culled from the twelve assessments contributed by these four teachers. Furthermore, these prompts were sub-questions within the entire task. In fact, such questions were few in comparison to prompts

focusing on knowledge and facts. By comparison, the *more aligned* teachers designed assessments that addressed a range of AIW criteria (an example given was Totoro's fieldwork assessment in Figure 5.1).

The *moderately aligned* teachers frequently adopted assessment goals that focused on preparing their students for examinations and tests. Therefore, they more often referred to assessment goals that were heavily skewed towards equipping lower secondary students with skills to do well in the upper secondary assessments, or to pass the end-of-the-year tests. They discussed attending to these test preparation skills more often than they spoke of longer term goals like preparing students for life beyond school. To this end, their assessment goals were quantitative in nature, focused on ensuring that students could accumulate marks. During all three interviews, Jiajia explained that she focused on higher-order thinking skills because she had to prepare her students for the upper secondary examinations. She noticed that in the past, the prompts in the national examinations "just required you to regurgitate the processes." However, the questions today "actually require some thinking." And so, teachers had to prepare their students to tackle such questions in order to "score well."

Similarly, Amanda pointed out that "why do we all do this? Marks!" She added that marks provided information for teachers "[by] placing students where they need to be." Miki also had quantitative goals for assessment because it is "a form of measurement of students' ability [and] for students to gauge where they stand." As a result, her goals were for her students to "answer [her] questions correctly." Since assessment was associated with the accumulation of marks or correct responses, these teachers were "very focused," because ultimately, they "have to prepare [students] for national exams" (Jiajia). Because of these

concerns, "assessment" was strongly related to "testing how much a student knows and remembers about a topic" (Margaret).

Given the goal of ensuring that students perform well in formal assessments, the *moderately aligned* teachers ensured that their assessments were designed following appropriate procedures and were focused on the skills assessed in the upper secondary examinations. For Margaret, assessment was first and foremost related to the "mark scheme," "multiple-choice questions," and "types of questions," indicating her familiarity with the procedures associated with test construction. Jiajia explained that paying attention to the "purposes, rationale, objectives" and "different modes of assessment" was necessary because "assessment drives everything." This concern was particularly true for teachers teaching upper secondary students, since they had to prepare these graduating classes for the GCE O- and N-level assessments.[22] Jiajia followed the examination criteria with great fidelity, and used them to determine what to teach. She explained,

> What I did for my upper sec classes is that I looked at the assessment—the format, what will be assessed: the content, the skills, everything. From there, I worked backwards to determine content-wise the areas to focus. Then skills-wise, which are the ones to assess them, or what to teach them.

This "backward design" (Wiggins & McTighe, 2005) process that begins with the upper secondary examinations determined the topics and skills that Jiajia and her colleagues taught at the lower secondary level. The importance attached to the upper secondary examinations meant that the lower secondary teachers needed to familiarize students—from an early stage—with the types of prompts as well high standards and rigor associated with these examinations. Hence in Amanda's school, there were efforts to standardize the test construction process. In the year prior

[22] These are aged-16 examinations called the General Certificate of Education "Ordinary" Level (G.C.E. O-level) that certify the end of secondary education. The results of these examinations are also used for the selection of students into post-secondary courses.

to this research study, Amanda's school began to coordinate all classroom assessments to ensure that similar assessment standards were applied across an entire year group. Prior to standardization and coordination, Amanda said "the marks difference is there, because I can set a simple test and everybody gets full marks. In another class, the teacher might be very harsh. Set a difficult one."

The evidence from these two examples suggests that assessments at the lower secondary level are conducted as mini-standardized tests in some schools. Although lower secondary is a non-high stakes level and that there is greater teacher autonomy at this level, the data indicate that the *moderately aligned* teachers did not use a wide variety of assessment formats or types. They followed the test blue-print strictly in order to familiarize students with examination-type formats. Ironically, this standardization at the school-level occurred during TSLN, a policy that envisioned devolving more curriculum autonomy to schools. The interview data revealed that teachers have organized themselves into teams to work on assessment, as compared to the past when each teacher prepared assessment tasks in isolation. In Amanda's school, collaboration ensured greater parity in the types of assessments administered to students across an entire year group, while in Jiajia's school, the team refocused teaching, learning and assessment to start preparing lower secondary students to meet the demands of the upper secondary high-stakes examinations.

Adhering to TLLM's intent to recognize the development of a learner, some schools moved away from relying on the midyear and end-of-year summative assessments as indicators of student learning. In these schools, there were numerous continual and semestral assessments

taking place over the academic year.[23] At the time of the research study, the schools were already using data from multiple assessments as indicators of student learning. As such, the end-of-the-year score for each student was based on the cumulative mark obtained from the continual and semestral assessments. Harry, a *more aligned* teacher, explained the computation

> We have the CA [continual assessment] and SA [semestral assessment]. So for CA, it's divided into different components. The common test is the highest weightage. I think 50 percent. Then you have your small test, or small project—you add up all together. A CA overall from January to March will comprise CA plus two other mini tests. So I may put this assignment as 20 percent, so from its converted x marks to 20 percent. Maybe another one, a small mini-test, will give you 100 percent.

Some schools have tried to align their practices with the TSLN vision and attempted to reduce the pressure and emphasis on high-stakes testing. In spite of this, there was a backwash. The midyear examinations at Miki's school were removed to provide lower secondary students with "more time to get used to the secondary environment," especially with their "having to do so many subjects."[24] Therefore, student learning was based on the combined results from class and common tests, as well as the end-of-the year results. However, this school-based assessment policy meant that each assessment now had higher stakes because the marks from all assessments were now computed into the final score. As a result of this computation of assessment scores, teachers conducted practice assessments, retests, and re-teaching. In the course of participating in the study, Miki conducted two retests. Margaret's third assessment was implemented in view of the "common test coming up" and so she wanted to "give them a bit of practice" on pie charts and graphs, because students tend to "get a bit stuck." In fact, Margaret's goal for this

[23] Continual assessments refer to the scores from assignments and class and common tests conducted throughout each of the ten-week terms, while semestral assessments are examinations that are held every six month, typically in May and late October, which are the midyear and end-of-year examinations respectively.
[24] In primary school, the examinable content subjects are English Language, Mother Tongue Language, Mathematics, and Science. However, when they transit to secondary school, the basket of content subjects increases. The examinable subjects are English Language, Mother Tongue Language, Mathematics, Science, Geography, History, and Literature.

assessment was for her students to "practice their exam techniques—how to answer questions, what's being looked for in the question."

The need to help students attain high marks for each assessment explains why teachers like Miki conducted retests. Doing away with the midyear examination meant that teachers had "more time to teach them, and [they] don't have to rush through the syllabus." However, the administrators realized that "[they] cannot just depend on the end-of-the-year mark because it is going to be very terrible for the kids" (Miki). Since most scores were now important, when students underperformed, retests were conducted. The aim was to help students attain higher marks, as Miki explained

> I will take the higher of the two to record for their CA. So for that class that did very badly in the first place, they will all be getting higher marks because I'm taking the higher of the two. A couple of them actually got lower. But because I'm taking the higher of the two, so it's to their benefit. Otherwise, it defeats the purpose of doing the second test.

Thus, on the one hand, the use of continual assessments documented student learning and development over one school year rather than evaluating students' learning based on one summative assessment. On the other hand, this practice pressurized both students and teachers. "If [a student] has not been studying or not following the lesson, she will be very stressed" (Miki). Teachers were further stressed because they had to "source around for questions, for pictures, and all this" (Miki) for the tests and retests. To this end, it is ironic that removing the midyear examination resulted in heightened stress and anxiety for both teachers and students because each classroom assessment now had higher stakes.

As a result of the accountability and administrative aspects of assessment, the *moderately aligned* teachers designed and implemented classroom assessments that mimicked mini-standardized examinations in both form and process. In terms of the process, the teachers worked

with colleagues to standardize the scheduling of assessments and to implement a common assessment across the year group. As to the form of the assessments, the three tasks Amanda submitted all adhered closely to the suggested format provided in the syllabus. While Jiajia contributed three teacher-designed worksheets, the formatting and wording of the prompts were closely aligned to the types of questions used in the upper secondary assessments. Likewise, Miki's three assessments resembled the format outlined in the syllabus document. Compared to the *more aligned* teachers, there was little or no variety in the assessment format and types submitted by the *moderately aligned* teachers. While the assessments adhered to the recommended test blueprint in the syllabus, the AIW scores showed that the assessment prompts were not always aligned to the Assessment Objectives. From the way these four teachers described the planning and implementation of assessments in their schools, it is evident that standardized practices occur pervasively and persistently, despite the fact that lower secondary education is supposed to be low-stakes, and that there is autonomy for more customized teaching and assessment at this level.

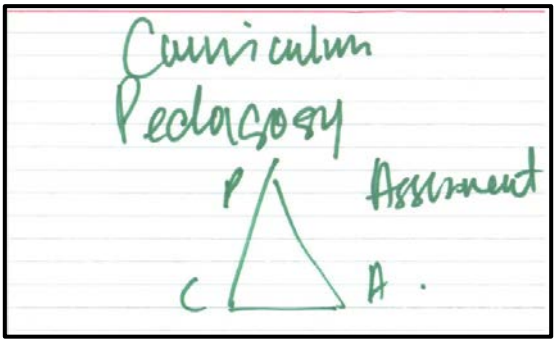
Formative assessment.

Similar to the *more aligned* teachers, the *moderately aligned* teachers adopted formative assessment strategies and made formative use of assessment information. The difference was that the *moderately aligned* teachers used formative assessment strategies that focused only on ensuring that students knew what the right answers were. In addition, they used fewer formative assessment strategies as compared to the *more aligned* teachers.

Formative use of assessment information. The *moderately aligned* teachers saw assessment as integral to classroom teaching and learning. Assessments were a means for Miki to "check on [her] students' understandings." Assessments enabled Margaret to find out "how

much students have grasped" of the topic. Based on students' work and their responses in class, formal and informal assessment provided information for Amanda to "make improvements" to her teaching. For Jiajia, assessment was "this whole cycle" (see her diagrammatic representation of her conception of curriculum, pedagogy and assessment in Figure 6.4) that included "assessment for learning" and "assessment of learning." In her view, "assessment for learning" required teachers to be "more responsive, more flexible and more prompt" in their teaching. In order to do this well, she believed that teachers "need to plan way ahead" and "to make it more structured." As "assessment for learning was a very weak part" of her practice, Jiajia tried to assess her students more frequently rather than wait for the results of their common tests results, which might be "too late" for formative assessment.

Figure 6.4
Jiajia's view of assessment, curriculum and teaching



Compared to the *more aligned* teachers whose goals were to find out what students knew and could do, the *moderately aligned* teachers aimed to examine *if* students knew facts and concepts. When discussing "what counts as student learning," the *moderately aligned* teachers focused on the marks the class obtained, or the percentage of the class scoring A-grades. The exception was Margaret who spoke about the "interesting" answers her students produced. She focused on what they could do, rather than how much they achieved. The two quotes from Miki

and Jiajia clearly illustrated their focus on the quantity of correct responses rather than the quality of those responses.

> At the end of the lesson, when I ask them questions, who are the ones who are able to answer these questions. These are the ones whom I thought, they can absorb the information very fast (Jiajia).

> If she has learnt from me, that means she is able to tell me, if I ask her a question regarding that, she is able to answer my question correctly and describe, let's say, the formation of this feature (Miki).

Adopting such views, the *moderately aligned* teachers focused on the number of students who managed to produce the correct answer. Therefore, when interpreting student performance, these teachers focused on the marks since marks are associated with correct responses. In response to the interview prompt, "How did the class perform?" these teachers used language that focused on the number or percentage of students who passed, and the number or percentage of students who scored distinctions (or A-grades). These teachers provided this pattern of responses for both formal (e.g., class and common tests), or informal assessments (e.g., class worksheets). For instance, during Interview 1, Amanda said that she had "at least 75 percent distinction" which is a score of between 17.5 and 18 on 25 for the first assessment. In Interview 2, an indicator that Amanda's students performed well was that "all of them got distinctions except for one." And in the third assessment, the indicator of student learning for Amanda—once again—was that "most of them can still get a distinction." Miki also associated student learning with the marks they received for the tests. At the first interview, she said that of the 6 high-ability students, 5 received "perfect scores." And at the third interview, Miki's emphasis once again was the marks students scored.

> The results vary – for 1 class, 1 girl from the medium-ability group scored full marks, whereas 2 girls from the high-ability group failed the test. For the second class, 2 girls from the high-ability group scored 9/10 and 10/10 respectively, while 2 girls from the medium-ability class scored 9/10 and 8/10 respectively.

The *moderately aligned* teachers and their students placed heavy emphasis on marks. Because students have been socialized to focus on "marks," Jiajia said that they "become very 'kiasu' [colloquial for being afraid of losing out]. They only focus on marks." Her students would painstakingly dissect the marks allocated to each assessment that they were tasked to complete.

> If they see a six-mark question or a four-mark question, they ask, does it mean that we write four points? Does it mean we write six points? They were very eager to make sure they score full marks.

The *moderately aligned* teachers adopted a variety of strategies to analyze student learning. Some were very precise, and involved a careful compilation of the items students answered wrongly or correctly. Margaret created mental checklists of strengths and weaknesses her students demonstrated after completing a task. Miki employed several strategies to analyze assessment data. During the first interview, for the map reading task, she tabulated and created a checklist of the number of map reading skills that her students were able to do. She also noted the areas of weakness: for distance, "they don't know how to convert the cm to km" and for grid reference, "quite a few of them also made the mistake by quoting the Northings first, instead of the Eastings." Miki's approach indicated that she looked beyond correct answers; she also focusing on the nature of the misconceptions and misunderstandings.

Another approach that teachers used was to identify the skills at which students were weak. When comparing the students' performances in Assessments 2 and 3, Miki concluded that students did better in Assessment 2 because there were multiple-choice questions in that test, as compared to Assessment 3 which required them to produce lengthier responses, something her Secondary 1 students were not used to. Based on this information, her future assessments will

include prompts that require more extended writing, while simultaneously reducing the use of tasks that require short responses.

Margaret's practices differed slightly from the other *moderately aligned* teachers. Some of her practices were more similar to the *more aligned* teachers. She focused more on "how well the students understood the questions" and "how well the students were able to draw the route on the map." In differentiating the quality of work among her students, Margaret "decided by who gave [her] the more detailed answers really. Some of them actually had very different thinking." She observed that her students produced "very interesting answers—high quality answers."

For the map reading assessment, Margaret noted that a group of her students were weak at six-figure grid references. As a result, she reviewed this set of skills with them while assigning the rest of the class to work independently on other tasks. Following the review, she assigned another worksheet to the group to ascertain if they had understood the technique this time.

Formative assessment strategies. One purpose of formative assessment is to help students move from their existing level of achievement to the stipulated learning goals (Sadler, 1989). In view of this, the *moderately aligned* teachers would clarify and re-explain mistakes and misconceptions. However, given their emphasis on ensuring that students were well prepared for high-stakes assessments, they focused heavily on test taking strategies. Jiajia strongly believed that thinking could be taught structurally. To this end, she provided her students with a "3-step approach" to navigate data response questions. She would remind her students that "the first step is the overview, the second—what's obvious, and then third—support your answer." She would also "nag" them into remembering that *elaborate* meant "explain how, explain why." Similarly, Amanda would advise her students on test taking strategies, and she would frequently

remind students to "manage their time well." Comparatively, the *more aligned* teachers did not speak about test taking strategies but focused on teaching approaches to help students close the learning gap.

Constructivist assessment envisages that teachers will provide students with feedback, and with chances to rework the task. This would enable students to demonstrate that they had mastered the learning goals after feedback. However, for these teachers, after analyzing students' responses, mostly, there was no time to change instruction or provide students with the opportunity to redo the task or activity. Thus, the next time the students encountered the skill or content demand again would be at the midyear or end-of-the-year examination, if not at all. In this case, changes to instruction would be made for the next cohort of students. When discussing her students' performance on the topic, 'Rocks,' Amanda said that she would change the way she delivered the topic for the next cohort of students. She pointed out,

> Usually when I teach I will look at all this kind of mistakes, and for my next batch of students, I will say [that] this is the one where they will always make mistake. And I would just highlight again. So my future classes, I may have to put a little bit more emphasis on that.

In comparison to Amanda, Miki and Margaret found time to re-teach. While Miki conducted a whole-class review of the lessons, Margaret customized the review of topics. After the map reading assignment, Margaret separated her class into two groups. She re-taught one group and revised six-figure grid references with them. To check that this group grasped the skill, she gave them "extra worksheets and then I collect them at the end of class." For the more competent group, she would "give them extra worksheets, or tell them to go on and do a mind map of a different topic." Alternatively, they would work on "one thing they're struggling with at the moment or something they're weaker at, and read through it in the textbook." This was because she did not want them "sitting down there kind of nothing for them to do. They can do

other exercises or revisions as well." In this way, Margaret was able to help one group while not over-teaching the other. However, providing this differentiated feedback to a small group meant that Margaret had "two different workloads" in class that day, which, according to the teachers in the other two categories, would reduce the already tight curriculum time.

Miki also re-taught topics but her aim was for students to re-take the test. In the course of participating in this study, Miki conducted two retests because her students had performed abysmally during the first test. During Interview 1, she discussed a "retest" she was planning for the class, and in this repeat of the same test, she decided to make the test simpler by using multiple-choice questions instead of the open-response questions that she had used in the first test. Similarly, the third assessment that Miki submitted was also part of a test and retest approach. This time, she submitted the retest for the study. She conducted a retest because—once again—her students had performed poorly. She hoped the scores on the retest would ensure more presentable marks for the end-of-the-year cumulative grade. The prompts on the retest were based on her analysis of students' responses to the first test. Miki "purposely" made the second test "very similar to the first." She described the differences between the two tests:

> The first test, I asked them to define 'watershed' and 'drainage basin.' For the second test, I gave them the diagram to show the drainage basin, and they are supposed to label the 'drainage basin' to identify the 'watershed.' Though put it in a different way but still the same concept. [In] the second test, I added on one more, to ask them about the source of the river. For that, actually quite a lot of them were a bit confused between [the] source of the river, and the river delta, the mouth. Because the watershed was placed in a different way. And I asked more or less the same for Question 2, except that the question on 'dams' I re-phrased it to make it clear for them. The second question asked about river features, which is exactly the same as the previous test.

Miki's comments indicated that she revised the prompts in her retest based on students' comments on the first test. Some students had said that the prompts were unclear in the first test. Therefore, the retest is based on Miki's formative use of assessment information through

dialogue with her students. Following the review and retest, Miki's students scored better in the second attempt. The quality of work following the re-teaching was also better, and this explains why her overall AIW class mean had higher ratings (ranked 3 in Table 5.13).

Retests were opportunities for teachers to expose their students to different permutations of assessing the same concept. After the second test, Miki

> clarified to them about the diagram. I turned the paper around, and I have the sea facing down, and I said, "If this is given to you, would you be able to see it better?" They said, "Yes;" because their idea is the river flowing down. But [when] I turn it this way, they can't see. So I said, yes, you must be exposed to all this, because in the exam, we can put it in different ways.

From this, it is evident that Miki did not want her students to take a replica of the first test. By varying the position of the diagram, she was in fact assessing her students' ability to apply the concept from a different perspective. Miki's rationale was to let students know that in the examinations, the same concept can be assessed in different ways. In comparison to Totoro (*more aligned* teacher) who wanted her students to deal with ambiguity, Miki's approach was to prepare her students to be familiar with the different ways a topic can be assessed. This exposure would reduce the unfamiliarity when students eventually took the year end examination.

One of the benefits of formative assessment is that following the provision of feedback, there are impressive effect sizes in terms of student learning (Black & Wiliam, 1998a; Hattie & Timperley, 2007; Shute, 2008). However, the *moderately aligned* teachers were dismayed that despite repeated teaching and the meticulous pointing out of errors, there were content, skills, and concepts with which students continued to struggle. Miki observed after the second hydrology test that her students were still unable to "handle the question of the flipped water shed." In all her interviews, Amanda spoke of her students' failure at applying and transferring

their mathematical skills to data interpretation in geography. She lamented that "they just cannot link maths to geography." To this end, she said that she needed to make it explicit for them

> When I teach, especially topo map, when I talk about eastings, northings, I would say to them, "It's like a graph. You read the *X* and the *Y*. Just like you read the eastings and the northings. Eastings then northings." I do the connection for them.

The repetition of errors confounded the teachers. Why was it that after re-teaching, students were still unable to grapple with certain concepts and skills? Why did they continue to make the same mistakes? Margaret commented that her Secondary 2 students were still "unsure" as to how to work with data response questions. Even experienced teachers like Amanda and Miki, each with over ten years of experience, were puzzled. Jiajia was similarly perplexed that, by the middle of the school year, her Secondary 2 students were still struggling to master basic content and skills, despite feedback and retests. These teachers had devoted much time talking to students to elicit their misconceptions and to understand their thinking processes. For instance, when she returned the graded assessments to her students, Miki would always

> find out from the students what exactly happened—is it because you didn't learn well, you didn't study or is it really the questions are not clear. If they did not study, then it's not my teaching problem. But it's really they have no time or whatever. If it's concepts that they are not sure of, then I would have to think through and see how I can make concepts clearer the next time.

The *moderately aligned* teachers had two feedback approaches. The first approach was the method of the feedback and the second was the nature of feedback. In terms of the method of feedback given to lower secondary students, the teachers provided feedback to the entire class. This was the usual practice for class and common tests because they were pressed for time, given that they have just two 40-minute periods a week (see CPDD, 2005). Typically, the feedback given to the lower secondary classes was closed and specific, provided to ensure that students knew the correct answer, and to guarantee that "they will be prepared" for the end-of-the-year

examination (Amanda). After marking and analyzing student work, Jiajia's approach was to "go through with the class what's the common mistakes." She would identify the main areas of weakness and tell students "you shouldn't just give me the answer, which is just note down the name of the country. Give the data, support it with data from the source." Another approach she used was to write examples on the board or, "the answers in the soft copy so I flash it through the projector." She would go through the "strengths." More time is spent highlighting the "common mistakes to look out for" because she thinks "that's important. That's usually what we do." Similarly, Amanda would "go through the test and highlight the weaknesses." She said that

> When there is data, they never give data support. Having the direction words, they never pay attention to the direction words. When they are supposed to describe, they go and account. So I'd highlight that. And what are the key words. Then also look at the mark allocation, [it] guides how much you have to write. So mistakes that are repeated a few times, then it's alarming!

Amanda's comment illustrates that her feedback was not just about the type of student response, but also drew students' attention to test taking strategies.

The *moderately aligned* teachers provided feedback that came in the form of the correct answers expected for the test prompts. They provided feedback that converged towards the ideal or "correct" answers as required in the mark scheme, textbooks, or publisher notes. Typically, these teachers did not allow for alternative answers, unlike Totoro, who always ensured that her mark scheme was flexible enough to accept other perspectives. In addition to oral feedback provided to the entire class, precise and directed feedback was written on the pieces of graded student work. For the map reading assessments, the teachers would tell students to "check directions properly" (Miki) or "work on your six-figure grid references" (Margaret). Otherwise, the response either had a mark or a cross to indicate whether the answer was right or wrong. The teachers provided specific and directed answers for the map reading tasks because the prompts

require precise measurements, and students needed to know how to apply and use the procedures and processes.

As these teachers provided whole class feedback, there was minimal customization of feedback for different groups or individuals. One reason was because they taught many lower secondary classes. Miki taught four of seven Secondary 1 classes, each with about 40 students. Individualized feedback was not provided for lower secondary students unless there were "very glaring mistakes" (Amanda). The heavy teaching load also resulted in teachers providing only brief comments. To help the students focus on the errors in the structured questions, teachers would underline a sentence or a phrase and indicate, "how?" and "read question," (Amanda) or where elaboration was missing, they would write "description?" (Miki). Sometimes the teachers would want something more specific, "Which year are you referring to?" (Jiajia). Other times, they communicated the misconception, "Floodplains and levees are considered as 1 feature" (Miki). The approaches and strategies the *moderately aligned* teachers used to close the gap were to "tell" students about their weaknesses, and then have them practice to obtain the expected responses through corrections. These written comments did not indicate students' strengths, clarify misconceptions or identify ways to help students to improve.

At the other end of the continuum, some teachers provided vague feedback, typically to praise and motivate the learner. They wrote vague but encouraging comments like "try harder" (Jiajia) or "excellent" (Amanda). Sometimes the comments might be open, for instance, "if you have difficulties, come to me" (Jiajia) or probing, like "consider labor shortage?" (Amanda).

Some teachers provided feedback in kind to boost students' self-esteem or to acknowledge their efforts, especially if the task had been challenging. Amanda distributed candy and chocolates to students who performed well, and Jiajia amused her students with pictorial

stamps that complimented students with words like "big effort." Amanda rewarded her class with lollipops after the second assessment because "all of them got distinction except for one." In the end, she relented and provided a reward to this student as well. This treat motivated the student because "in this test, he worked a bit harder" (Amanda).

However, providing extrinsic rewards in this form is not feedback because it does not provide information about the task and students are not shown what to do or how to improve. In fact, such practices have a backwash effect because learners lose motivation once the reward is withdrawn (Hattie, 2009). Feedback that is given to boost morale and increase motivation does not inform them of what they have achieved nor guide them towards the next steps or strategies they can take (Black, et al., 2003b).

Role of the student

The *moderately aligned* teachers controlled formative assessment practices in the classroom. In the day-to-day busyness of the classroom, there was no room for student voice because both teachers and students had to be "very focused" and teachers "have to prepare [students] for national exams" (Jiajia). Similarly, Amanda preferred that teachers continue to play the dominant role in the classroom. She said that "even if it's presentation and rubrics, it has to be teacher!" As a result, students were relegated to a passive role, merely adopting, accepting, and adhering mechanistically and routinely to the strategies and comments provided by their teachers. These responses differed somewhat from those of the *more aligned* teachers who, from time to time, would attempt to provide space for students to play a more active role in learning, such as by asking them to identify what was wrong with their responses, and to suggest how they could improve.

Among the *moderately aligned* teachers, Jiajia and Miki would sometimes provide opportunities for a little student participation. These practices, however, differed from the open questioning and guiding that Harry and Totoro undertook. Jiajia's students were allowed to grade each other's responses. Typically these were responses to closed-ended questions like multiple-choice or short answer questions, where there was one indisputable, correct answer. For open-ended questions, Jiajia would get her students to grade their own work so that "they'd actually know where they've actually missed out on, and what they need to do well."

Ideally, some teachers would like their students to be more active in class. Miki spoke about her "rather passive" students who were "shy to raise their hands, shy to answer even if called upon." Occasionally, she would encourage her students to write or draw their response on the board. She observed that "those who are a bit more outgoing, the gung ho ones would run up and do, even though they are wrong." However, "there are a lot who are not very ready." One of Miki's goals would be to "have more of that kind of interaction" where students would want to volunteer responses during class discussions.

Summary

The *moderately aligned* teachers emphasized shorter term goals: their main concern was to ensure that their lower secondary students were well equipped with the skills and content knowledge required for the upper secondary high-stakes examinations. In general, the *moderately aligned* teachers' comments indicated that their assessment practices focused on preparing students for life and for test preparation. However, while their goals were about preparing students for life in the 21st century, their AIW scores and their comments suggest that these teachers employed assessment practices that were more aligned with test preparation and test scores. Although their assessment goals were aimed at developing students' higher-order

thinking skills, in practice, their assessments closely mirrored the formats used in the upper secondary high-stakes assessments. Within their schools, these teachers and their colleagues established procedures to streamline the lower and upper secondary syllabuses so that they could start preparing students for upper secondary assessments.

Given that their goals were to prepare students for upper secondary examinations, the *moderately aligned* teachers focused on teaching students strategies for performing well on different kinds of assessments (e.g., Jiajia's three-step approach to answering data response questions) or on exposing students to different permutations through which a concept could be assessed (e.g., Miki warned her students that teachers could assess the same concept in different ways). For formative assessment, these teachers ensured that students knew the correct answer. They did not provide opportunities for the students to figure out the errors themselves. Therefore, classroom assessments were controlled and driven by the *moderately aligned* teachers, with little opportunity for students to participate actively.

Less aligned **teachers and classroom assessment**

The two *less aligned* teachers, Maryanne and James, adopted assessment practices that deviated most from the TSLN intent. Their comments and the types of assessments they meted out to their students resonated most strongly with an earlier phase in Singapore's education journey, the *efficiency-driven* phase in which objective tests were used to stream (or track) students into different ability courses (OECD, 2011). These teachers' practices had characteristics of behaviorist learning theories, especially in the ways they approached knowledge and learning. For instance, they focused on test-teach-test (Shepard, 2000, p. 6) and viewed knowledge as being acquired (Sfard, 1998).

Assessment and learning goals

These *less aligned* teachers expended their efforts on ensuring that their students obtained good marks and grades during common tests and the end-of-the-year examinations. As a result of this learning goal, they conducted remedial lessons, re-taught topics, issued numerous practice papers, and implemented retests. They had short-term assessment goals that were focused on getting their students through the school year with good grades.

Based on these learning and assessment goals, it was not surprising that the assessments that James and Maryanne submitted closely mirrored the format suggested in the test blueprint recommended in the syllabus (shown in Table 5.4). These two teachers did not use other forms of assessments during the study period, even though the syllabus recommended that teachers use of a variety of assessments. The assessments used by the *less aligned* teachers had the least variation in terms of the format and skills assessed. In fact, they focused only on assessing knowledge and facts, as indicated in their low AIW scores. While they adhered to the test blueprint recommended in the syllabus, the nature of the prompts was not aligned with the Assessment Objectives in that these teachers only assessed knowledge, and paid little attention to the skills of Critical Understanding and Constructing Explanation (AO2) and Interpreting and Evaluating Geographical Data (AO3).[25] Interestingly, of the eight teachers, these two teachers were the only ones who submitted worksheets and activity sheets that were created by a vendor and not by the teachers themselves. The other six teachers submitted tasks that were teacher-created, and even if they had referred to a textbook or activity book, they modified and adapted the tasks to the needs and abilities of their students.

[25] See footnote 1 of this chapter.

As shown in Figure 5.2, Maryanne's assessment following the visit to the Botanic Gardens was comprised of prompts designed to elicit students' knowledge of discrete facts about the tropical rainforest. Similarly, Maryanne's third assessment merely required students to refer to specific pages in the textbook and "draw or/and label diagrams to describe the features of a river system." Her goal was for students to "internalize what these pictures are" and one way of achieving this goal was for students to "draw or redraw them." When compared with the *Communication* assignment used by Margaret (*moderately aligned* teacher) that required students to draw their conception of the term, Maryanne's task only asked students to reproduce diagrams and textbook definitions. Although the two tasks required students to demonstrate their learning by presenting a graphic, Margaret's task made higher demands of students because students first had to explain "Communication" in their own words and then design a graphic that best represented their conception of the term. Comparatively, Maryanne's task merely occupied students in "busy" work (Hayes, et al., 2006, p. 114), as she only required them to copy definitions and reproduce textbook diagrams. The tasks Maryanne assigned her students merely required them to work on repetitive procedures and offered them little intellectual challenge.

Finally, unlike the *more* and *moderately aligned* teachers, the *less aligned* teachers did not include geographical attitudes and values in their learning goals and none of their assessment prompts addressed these areas. For example, while Harry (*more aligned* teacher) asked his students to reflect on how they would apply what they learned in class to their lives, and Jiajia (*moderately aligned* teacher) tasked her students with commenting on how humans can protect the environment, such prompts and tasks did not appear in the assessments that the *less aligned* teachers submitted. In fact, the *less aligned* teachers' assessments focused on reproducing content. This was similar to the types of prompts Maryanne provided for her fieldtrip assessment.

Likewise, for James, the goal of assessment was simple, "what I want is for this kid to just remember, see the question, and vomit everything out," he said. "But," he continued, he also wanted to "make sure, whatever their vomit out—their answers—they try to answer the question."

The *less aligned* teachers subjected their students to many forms of test preparation. To ensure that students were test ready, James had "gotten some revision books, and zapped the summaries of the revision books." His strategy was to expose his students to "permutations of this type of question, so that they are not caught unaware." Assessment in James' school was "still very much a drilling kind of thing. The more questions you set them, the more familiar they are with the answers, the answer formats, the better they'll do." During the interview, James pointed out that his students did very well on the third assessment because they had previously attempted and practiced on different versions of the same prompts. An example of how James exposed his students to different prompts that assessed permutations of the same concept is shown in Figure 6.5. By contrast, among the four *moderately aligned* teachers, only Miki conducted retests.

Figure 6.5
Repeated prompts[26]

- What do you think are the reasons responsible for water constraints?
- Explain how the following can cause water constraint?

  (a) Rapid population growth and increased human activities
  (b) Polluted water in rivers and lakes
  (c) Change in global climate

The *less aligned* teachers had assessment goals that were associated with the acquisition and improvement of marks. Conducting retests was one way to increase marks. Another way was to give students pep-talks. James described strategies he used to "motivate" his students to

[26] These prompts were extracted from the second assignment James submitted.

study harder to accumulate marks. He impressed on them that mathematically it was easy to accumulate sufficient marks to get an A-grade on his Secondary 2 geography midyear examination. Prior to the midyear paper in May, he informed them that "the paper is only upon 60. You only need 42 marks to get a distinction. It's the easiest distinction you can get. Because if it's a hundred-mark paper, you will need to get 70 marks." In order to ensure that his students did well as a class, James cajoled his higher-ability students to "do better to pull their weaker classmates up." At the same time, he cajoled the weaker students not to "drag the whole class down." James's strategy worked because for the midyear examination, his top scoring student had 51 marks out of 60, and there were just "8, 9 failures in the class." However, this approach only serves to deepen the mark-focused assessment practices that TLLM is encouraging teachers to emphasize less.

Formative assessment

Formative assessment strategies used by the *less aligned* teachers were dedicated to ensuring that students had model answers. When making formative use of assessment information, these two teachers focused predominantly on what students were not able to demonstrate. Therefore, they saw their students from a deficit perspective, as compared to the *more aligned* teachers and some *moderately aligned* teachers who valued what students demonstrated they could do and used the assessment data to stretch students further.

Formative use of assessment information. For the *less aligned* teachers, student learning was associated with the number of correct responses students produce. Therefore, higher marks, as indicated by the number of correct responses, meant that students had studied their textbooks and notes. When analyzing student work, an indicator that his students had not learned was that the student "has totally no content knowledge. That's why he got 10 upon

60. … Nothing for him to write." In another assessment, James interpreted that his students performed "exceptionally well" on the third assessment. This was because in a class of more than 30 students, "there were around 27 'A's. And there were only 4 failures from the class."

When examining student performance, these two teachers related the marks obtained to the time and effort students devoted to memorizing. Commenting on the first assessment she contributed, Maryanne said that the aspects of that assessment that worked well were that the "Express [course] students [succeeded] through sheer memory work, recall major features, major ideas." When her students failed to do well in her assessments, it was because they "could not remember a lot of locations, especially the names of fold mountains." Ultimately, Maryanne was concerned with "how to make [her] test in such a way to encourage [her] students to retain content longer." James also focused on the amount that students had learned. He pointed out that students who performed abysmally were lacking in the quantity of content. When his students "aced" the third assessment, he attributed this to their being able to "easily memorize important bits." Another reason for their good performance was exposure to numerous "similar question types, similar concepts in the practice papers." Thus, working on mock papers was "a good practice for them."

Formative assessment strategies. When students did not do well, the implication for these two teachers was to re-teach or re-explain, and then retest if there were time. Unlike the *more* teachers, these two teachers did not talk to students to elicit and understand their misconceptions.

Maryanne and James adopted different formative assessment approaches after they analyzed the data from students' work. For James, there were few opportunities for students to demonstrate the change in learning based on the feedback. He pointed out that even with

assessment information, he was unable to make changes to his teaching plans, especially for current students. This was because "there is a scheme of work—it's already planned." As a result, he had no choice but "still had to proceed with the set assignments, even though [students] might not be ready for it." Thus, unless the same questions or prompts were used in the end-of-the-year exam, the students would not have a second chance to attempt the task to show that they corrected their conceptual misunderstandings and had closed the gap in their learning. This differed from Harry's (*more aligned* teacher) practice of providing his students with feedback and having them re-draft their work.

Comparatively, Maryanne, like Miki, found the time to reteach and retest her students. On the day of her first interview, Maryanne was implementing a retest of the assessment she was contributing. She had to re-administer the test because her students "failed" very badly. Before the retest, she re-taught the topic. One approach she used was to

> start off by showing them something intriguing. For example, there was this time when I showed them pictures of Mount Vesuvius and the city of Pompeii. They started asking a lot of questions: why is it so, why is it like that. That would be a better way instead of just giving them a map of the world.

While Maryanne's strategy was to find a different way to teach the topic and to pique students' interest in the topic, her comment did not show how she diagnosed students' errors or attended to their misconceptions.

Like the other teachers in the study, the way these two teachers provided feedback varied by grade level. Lower secondary students received whole class feedback while upper secondary students received a mix of whole class and individualized feedback. The aim of feedback was to help students obtain higher marks and to reduce errors in the end-of-year examination.

Whole class feedback was an efficient approach to reach out to students, given that they had time constraints. Copying answers from the board was one productive way to transmit

correct responses. In fact, after the third assessment, James insisted that his students copy model responses to the definition of *water catchmen*t, *desalination*, and *potable water*, even though some of them had received the full score. His rationale was that students' responses were too lengthy, and in comparison, "these are model answers, concise, straight to the point." In doing this, however, James did not explain misconceptions or point out gaps in the learning. He merely wanted to ensure that students had religiously copied the correct answers. He impressed on them that the end-of-the-year examination would be more demanding than their common tests, and so, "they were quite serious in copying down the model answers." Similarly, Maryanne's approach to rectifying misconceptions was to "tell" students the correct answers.

James used more "verbal feedback" than written feedback. At the time of the study, James taught over 200 lower secondary level students in seven classes. Since he taught the entire level, and had to "mark [the] entire cohort," he did not "really have time to pen down little comments and all that stuff." There was "no way for [James] to explain to [the lower secondary students] that they need to elaborate, elaborate, elaborate." When pressed for time, he would draw "inverted 'v's" on his students' responses to indicate that there are facts and ideas missing. This was the practice he adopted with his lower secondary classes. In the same way, Maryanne provided feedback "to the whole class." This was because when scoring the papers, "[she] could get a sense …like how well have [students] done for a particular test." As a result, she would provide "general remarks for how well they have done for certain components." Though at times, as with a student who did not complete more than half of the paper, she would "speak with them individually."

Comparatively, more individualized and customized feedback was only provided for the upper secondary students. James said that with the graduating classes, he was able to provide "written feedback." He said that

> for upper sec, I'm catering my feedback to each and every one of them. So every piece of work that I mark, I will write something, and I will return the papers to them individually, and with the model answers. And I will point out, I will go question by question to point out what went wrong, what needs to be done.

James found time to meet with his upper secondary students individually because he only had to talk to "twenty students at a time, for one and a half hours." So when they are doing the given work for that day, he "can hand in their previous work and give them feedback." One reason for the variation in the formative assessment practices James used was due to the nature of the students. He believed that open discussions of questions following a test would not work in all his classrooms. Based on his knowledge of his students, James said that discussions might be more animated among his upper secondary classes. However, the lower secondary level "class will be very quiet…. Only a few will answer…the rest will remain quiet."

To help students score, these two teachers had tried-and-tested strategies. Maryanne's formative assessment strategies were based on "practice makes perfect." She made her students rehearse the strategies or practice answering more questions with "stimulus, pictures, and graphic sources." Like Jiajia—a *moderately aligned* teacher who gave her students test taking strategies—Maryanne always "told" her students "how to approach a particular question." James also ensured his students were cognizant of test-taking strategies. When his students were unsure of the right responses to the multiple-choice questions, James "told them about the elimination method, try to get them to understand that there is always a correct answer somewhere in there, so strike off all the wrong answers first."

Finally, the nature of the feedback varied with the purpose of assessing. For in-class assignments, these two teachers provided little or no feedback. For instance, James did not go over the in-class task item by item, but provided general comments. He merely returned the student responses through the "subject reps and they will just pass the papers around. No feedback will be given, unless they did well. Then I just scribble a 'good' or an 'excellent.' Other than that, there will be quite little feedback." In comparison, for class tests there will be some written or oral feedback. Typically, James would provide more verbal feedback because he would "pass them their papers individually. So when I pass it to them, I will tell them, 'good job,' 'you can do better,' or 'a bit more.'" Such feedback is vague and focuses specifically on boosting morale. It does not provide students with a specific means to improve their learning.

Role of the student

There was little opportunity for student ownership or participation in these two teachers' classrooms. Neither teacher spoke about creating room for student autonomy. These two teachers relegated their students to passive roles. Maryanne stressed during each interview that her students were "empty vessels that [she had] to fill" while James called his students "robots." Neither teacher was enthusiastic about using opening questions as a formative assessment approach. James pointed out that because he "[treats] them like robots, they behave like robots."

Summary

The *less aligned* teachers had assessment goals which focused on accumulating marks and passing examinations. Their learning goals were to ensure that students learned basic facts and concepts by heart. Their assessment tasks used formats that adhered to the test blue print suggested in the syllabus for examinations. While the syllabus encouraged teachers to use a variety of assessments to capture student learning, the *less aligned* teachers did not vary their

assessments format. Consequently, their assessments did not meet the Assessment Objectives stated in the syllabus. In fact, the low AIW scores assigned to the assessments they submitted attest to the fact that their assessments did not attend to higher-order thinking skills.

The formative assessment practices of the *less aligned* teachers involved their drumming in facts through routine practice, repetition and re-teaching. There was little autonomy accorded to students. There were no instances in which these teachers provided opportunities for their lower secondary students to engage in making sense of their work. Feedback was provided *en masse* to the whole class after common tests, with little or no customized feedback to meet individual needs.

Comparison across the three teacher groups

The TSLN vision had time over a fifteen-year period to take root in Singaporean classrooms. At the policy level, the realization of the TSLN intent was supported by reviews of the syllabuses and national examinations to guide the use of teaching and assessment practices that would enable students to be prepared for life beyond school. Based on the micro or classroom level data, how aligned are the eight teachers' classroom assessment practices with the policy intent? This section compares the classroom assessment practices across the three groups of *more*, *moderately*, and *less aligned* teachers. Drawing on the three features of constructivist assessment—assessment and learning goals, formative assessment, and role of the student (shown in Figure 6.2)—the comparison of classroom assessment practices used by the three groups of teachers suggests some similarities and differences in the approaches to classroom assessment. Of the three features of constructivist assessment, the groups of teachers differed most in their assessment and learning goals as some teachers focused on preparing students for life outside school while other teachers put their emphasis solely on passing examinations with

good grades. The practices of teachers in the groups were most similar for the feature, *Role of the Student* in the sense that the majority of teachers did not provide opportunities for students to engage in peer or self-assessment or to play an active role in constructing assessment tasks and criteria. Table 6.2 presents a summary of the practices.

Table 6.2
Comparison of classroom assessment practices across the three categories of teachers

Feature	More aligned	Moderately aligned	Less aligned
Assessment and learning goals	• Focus on preparing for skills beyond school (long term) • Emphasize thinking skills, and ability to deal with ambiguities • Focus on content as well as application to the real world (as seen in AIW scores)	• Focus on preparing lower secondary students for upper secondary examinations (medium term) • Emphasize thinking skills that are aligned with upper secondary examinations • Focus on content and skills that are aligned with what will be tested at upper secondary level	• Focus on ensuring students pass the end of year examinations with good grades (short term) • Do not mention thinking skills specifically • Focus on content needed for end-of-year examination
Formative assessment	• Emphasize the quality of student response to identify gaps in understanding • Discuss the quality of responses as indicators of student learning / highlight what students show they are able to do • Identify strengths and weaknesses in student work • Sometimes prompt and guide students to figure out why the responses were incorrect • Do not ask students to concentrate on	• Emphasize marks as indicators of student learning • Analyze the quality of student response briefly with more emphasis on marks and grades • Discuss student work from a "cannot do" perspective • Provide correct responses in a lecture-style approach • Go over strategies for answering questions (especially data response questions)	• Emphasize marks as the indicator of student learning. The more marks the better • Do not focus on the quality of student responses • Discuss student work from a "cannot do" perspective. Highlight what answers are wrong but do not provide reasons or comments as to why responses are wrong • Focus on ensuring correct answers are provided • Ensure students copy

Feature	More aligned	Moderately aligned	Less aligned
	model answers and examination strategies		model answers
Role of student	• Do not provide room for peer and self-assessment • Do not involve students in joint construction of expectations and assessment criteria	• Do not provide room for peer and self-assessment • Do not involve students in joint construction of expectations and assessment criteria	• Do not provide room for peer and self-assessment • Do not involve students in joint construction of expectations and assessment criteria

Assessment and learning goals

Mainly, the teachers' assessment and learning goals did not reflect the TSLN intent. Of eight teachers, only four had assessment and learning goals that echoed the policy vision. The assessment and learning goals represent what the teachers intended their students to achieve. These goals also reflect the teachers' expectations, an aspect of constructivist assessment, as shown in Figure 6.1. Across the three groups, the *more aligned* teachers had goals that were most likely to focus on the long-term development of their students. While they also helped students prepare for the end of the year examinations, they did not make this their main goal. Instead, they wanted to assess attitudes and skills that would prepare students for life outside of school. Consequently they used a variety of assessment formats and types, and assessed a range of content and skills, including reflection. Comparatively, the *moderately* and *less aligned* teachers were more likely to emphasize examination skills and preparation. The *moderately aligned* teachers were more likely to focus on equipping students with skills and knowledge years before the upper secondary examinations while the *less aligned* teachers were preoccupied in testing and retesting their students to make them ready for the end-of-the-year examinations. These teachers used assessments that mirrored the examination test blue print. There was also no variation in the assessment formats used.

The teachers' learning goals were manifested in the types of assessments the teachers designed and implemented in the classroom. The *more aligned* teachers used the largest range of assessment types, reflecting the suggestions in the syllabus. As their goals were not merely for their students to pass examinations, their assessments used prompts and tasks that mirrored the AIW criteria for higher-order skills. Comparatively, because the *moderately* and *less aligned* teachers emphasized preparing students to pass examinations, the assessments they used adhered more closely to the test blueprint recommended in the syllabus document. In spite of this, the nature of the prompts did not reflect the higher-order skills reflected in the syllabus. To reduce the reliance on using just one mode of assessment, the syllabus urged teachers to adopt a wide range of practices so that more skills could be assessed. However, as reflected in the AIW scores presented in Chapter 5, only Harry and Totoro—the two more aligned teachers—assessed the types of higher-order skills as envisioned in TSLN, and as reflected in the geography Assessment Objectives. Apart from Harry and Totoro, the other teachers did not receive AIW scores that were above the scale midpoint for one or more of the assessments they submitted.

The teachers' learning goals and the nature of assessments used were closely associated with the teachers' expectations of their students. Mirroring closely TSLN's focus on recognizing the potential in every child, the *more aligned* teachers saw their students as being in a stage of development, a view consistent with constructivist theories (James, 2006; Shepard, 2000). In their view, each assessment merely provided a status report of the students' current level of performance. What was important, then, was not the current performance, but what students were capable of doing by themselves following assistance and support, but with the eventual aim of succeeding without any help. This view of student learning was aligned with the zone of proximal development (Vygotsky, 1978) in which teachers first provide scaffolds to students so

that they are able to complete tasks with some help, and then remove this support after learners demonstrates their ability to manage independently.

Comparatively, the *moderately* and *less aligned* teachers viewed students as having fixed abilities (Black, et al., 2003b; James, 2006) or saw them from a deficit perspective. These views of learning resonate more firmly with TSLN's preceding phase in which students were placed onto different courses depending on their test scores. Guided by these views of student learning, the *moderately* and *less aligned* teachers used assessments that focused on assessing basic facts. They also relied on external stimuli such as giving chocolates in order to motivate their students.

Formative assessment

Formative assessment practices are aligned to TSLN's goal of developing lifelong learners. In particular, formative assessment strategies prepare students to gradually take ownership of learning by guiding them to make sense of their work (Black & Wiliam, 2012b; Sadler, 1989, 1998). In the spirit of formative assessment, student learning as demonstrated in the work that students complete provides feedback to teachers who then interpret and make decisions to guide the next steps in teaching and learning (Black & Wiliam, 2009; Thompson & Wiliam, 2008). Formative assessment therefore includes the use of strategies such as feedback to guide and demonstrate to students how they can effect improvement (Sadler, 1989). Teachers need to use the assessment data to help students incorporate and integrate this information so that they can improve their learning (Black & Wiliam, 2009; Sadler, 1989). Providing feedback during formative assessment, therefore, suggests that there should be opportunities for students to revise and re-work tasks following feedback and guidance from the teacher. For this reason, feedback involves teachers providing students with information that is specifically related to the task or process of learning. This process is designed fill the gap between the intended

curriculum tasks and goals of what students are meant to understand on the one hand, and what students have demonstrated they have understood on the other (Black & Wiliam, 2012b; Hattie & Timperley, 2007; Sadler, 1989). Therefore, drafting and reviewing work based on feedback are opportunities for students to learn from their mistakes and to engage in continuous learning and improvement.

In general, the data from the interviews provide evidence that due to time constraints there were few opportunities for students to revise their work based on feedback. While Harry spoke of providing feedback and then asking his students to re-draft their work, the instances in on which he did so were limited. As a result, there were few occasions where students could redo their work following specific feedback to help them improve. This was the case even for the *more aligned* teachers.

One reason why practices diverged from the policy intent was that teachers needed to push ahead with the next portion of the syllabus. As the *less* and *moderately aligned* teachers mentioned, the large number of classes such as the ones that James and Miki taught constrained them from providing time for students to re-draft their work—there was too much reading and marking. Furthermore, teachers like James, Amanda, and Miki did not devote time to customize and individualize feedback for the lower secondary students. By comparison, these teachers only provided customized feedback for the upper secondary students because class sizes were smaller and the students were preparing for high-stakes examinations.

Formative assessment provides information that feeds into and informs teaching and instruction (Lambert & Lines, 2000). The analysis of how these teachers enacted formative assessment indicated that by and large they used assessment data as feedback to feed forward to their teaching and lesson planning. This was evident in the way teachers like Miki and Totoro

made meticulous lists and categories of student errors in order to identify the areas in which the conceptual misunderstandings could be rectified. This compilation provided a checklist of what students could and could not do. The formative use of assessment data pointed teachers to the areas students were struggling with, and was supported by approaches to address these weaknesses.

However, the way the teachers used formative assessment practices diverged from the TSLN intent. In particular, teachers did a lot of "telling," a practice that TLLM was exhorting teachers to do less of. For the majority of the teachers, the feedback they provided took the form of "banking" (Freire, 2000) in which teachers deposited the correct answers through mini-lectures or re-teaching, and expected their students to retain and remember these responses or advice. There were few opportunities for students to ask or to understand where they had gone wrong. Other than Harry, who provided feedback that highlighted his students' strengths and provided suggestions for improvement, teachers merely marked and provided short responses that included the right answer or a short remark to boost confidence. Even among the *more aligned* teachers, there was little evidence of teachers "guiding, facilitating, and modeling" as envisaged in the TLLM tenets (MOE [Bluesky], 2005). Although Totoro sometimes asked her students to discuss and examine where they had gone wrong, and to work out the responses themselves, she only provided this opportunity for her higher-ability classes. In response to their analyses of students' errors, what the *moderately* and *less aligned* teachers mainly did was to remind their students again and again about the content and skills as well as examination strategies that they needed to master for the end-of-the-year examinations or the upper secondary examinations.

Role of the student

In their recent discussions of formative assessment, Black and Wiliam (2009) argue that classroom assessment information should be actively interpreted and used by teachers, learners and their peers. This suggests that students must participate more in the assessment and learning process and be given the opportunity to actively make sense of their learning. The TLLM tenet which calls for teachers to do more "guiding, facilitating, and modeling" and less "telling," resonates with the vision of the student's role in constructivist learning.

Across the three groups of teachers, there was minimal evidence of an active or student-led role in the classroom. There was limited evidence of peer and self-assessment. Nor did the teachers give students the opportunity to engage in decision making processes such as developing a rubric or initiating assessment tasks. The teachers rarely sought to understand students' perceptions and thoughts. Even though the *more aligned* teachers wanted their students to be able to apply their learning to real world tasks, their view of the "real world," were largely conceptualized from an adult's perspective. The closest to providing students with decision-making opportunities was to provide them with a choice of topics to work on.

Of the eight teachers, only Totoro and Harry spoke of asking their students to figure out their misconceptions and errors. While the occasions were not frequent and did not always apply to all students, these two teachers valued such practices. Their perspectives were consistent with the academic discourse that urges more direct student involvement in educational reform (e.g., Levin, 2000; McQuillan, 2005; Mitra, 2009; Rudduck, 2007). Furthermore, decision-making opportunities such as that provided in Totoro's fieldwork assessment enable students to have "real learning" (Rudduck, 2007, p. 591). These activities prompt students to think and make decisions, rather than to rely on their teachers to provide answers. In addition, she provided

timely feedback when she led her students to reflect on the discussion and interview process while on the bus back to school. Rather than tell her students where and how the interview process could be improved, she elicited responses from them, and encouraged them to suggest alternatives. Totoro's practice of encouraging students to figure out answers for themselves is consistent with Sadler's (1989, 1998) conception of effective feedback, as she provided space and time for a more participatory student role in teaching, learning, and assessment. Her approach enabled her students to take ownership of learning as they constructed their learning through active sense-making. Totoro's practice is also aligned with the TLLM goal of engaged learning, because when students feel empowered in school, they are less disengaged and more accepting of the school's programs (McQuillan, 2005). Unfortunately, such practices were mostly confined to the two *more aligned* teachers.

Summary

Of the three constructivist assessment features, there was the most divergence among the three groups in relation to *Assessment and learning goals*. For this feature, the *more aligned* teachers' comments and practices were closest to the TSLN vision and TLLM tenets in spirit and in practice. The teachers spoke passionately about preparing their students for life outside of school, and this meant equipping students with higher-order thinking skills. As shown in the assessments they submitted, teachers provided opportunities for students to apply what they learned to real world contexts. There were more diverse practices among the *moderately aligned* teachers, with some (i.e., Jiajia and Margaret) adopting practices that were similar to the *more aligned* teachers, and some (i.e., Amanda and Miki) enacting practices that were closer to the *less aligned* teachers. The *moderately aligned* teachers had practices that were less similar to the TSLN vision in spirit because their assessment and learning goals were geared towards preparing

their students for the upper secondary high-stakes examinations. Thus, any higher-order skills that they emphasized were those that would be examined in the upper secondary syllabuses. The *less aligned* teachers enacted classroom assessment approaches that were least like the TSLN vision in spirit and in practice. From the analyses of their interview comments and of the assessments they submitted, it was clear that their emphases for assessment and learning were test preparation and content reproduction.

In relation to *Formative assessment*, there were differences among the three groups of teachers. The *more aligned* teachers spoke of building on what students' showed they could do and had learned, in comparison to the *moderately* and *less aligned* teachers, who focused on what students were unable to do. These latter two groups of teachers also used marks and grades as indicators of student learning. To this end, formative assessment for the *moderately* and *less aligned* teachers involved their telling students what the correct responses were, and equipping students with test taking strategies. In comparison, the *more aligned* teachers analyzed the quality of students' responses, and identified strengths that they could build on. Where possible, the *more aligned* teachers would avoid telling students the correct responses, but would try to guide and probe their students towards identifying and making sense of what went wrong in their responses. In this respect, their practices closely mirrored the TLLM tenet, "more guiding, modeling, and facilitating" and "less telling" (MOE [Bluesky], 2005). These teachers focused on helping students to learn.

The practices of the three groups of teachers converged in the third feature, *Role of student*. None of the three groups of teachers provided opportunities for student ownership in the learning process or involved students in peer or self-assessment.

The comparison among the three groups of teachers shows that, on the whole, the eight teachers' classroom assessment practices did not converge towards assessing the skills envisioned in the TSLN vision. While the *more aligned* teachers enacted practices that resonated with the policy, at the end of the day, their practices were a combination of constructivist and behaviorist assessments. Their practices captured elements of constructivist assessment because of their learning and assessment goals, and the way they practiced formative assessment. The types of assessments and prompts they presented to their students reflected alignment with the TSLN vision. However, because their assessment practices did not encourage more active student participation, this suggests that their approaches also diverged from constructivist assessment. The assessment practices of the *moderately* and *less aligned* teachers diverged substantially from the policy intent. The analyses of the assessments they submitted and of their interview comments suggest that their practices focused on preparing students for tests and examinations, rather than on student learning. This was evident in these teachers' emphases on marks and on making sure that their students knew examination strategies and right answers.

Factors influencing the nature and quality of classroom assessments used

The analyses and discussion of each teacher group, and the comparison of the three groups of teachers, suggest that, on the whole, some classroom assessment practices have been enacted that reflect the TSLN vision and the TLLM tenets. In particular, the *more aligned* teachers and some of the *moderately aligned* teachers enacted practices that were aligned with the TSLN vision. This was especially the case in terms of the learning goals through which they seek to develop and prepare their students for life after school. However, there appears to be less alignment with two other aspects of constructivist assessment. First, rather than using formative assessment to support learning, the majority of the teachers focused instead on dispensing correct

answers for students to memorize. Second, most of the teachers did not provide opportunities for students to be active members in the classroom.

In comparing the assessment tasks of the three groups of teachers, it is evident that there are some factors that support or hinder the teachers' design and use of authentic intellectual work. The data in the previous section analyzed using the constructivist assessment lens have shown that each group of teachers focused on different learning goals, and that these impacted the format and nature of prompts they employed in their assessments. The goals also affected the way the teachers worked with students and the manner in which they used formative assessment. The variation in the quality of the nature of the assessments was manifested in the different AIW scores the teachers received, as presented in Chapter 5. Following the comparison among the three groups of teachers in this chapter, this section draws on the qualitative data to explain the differences in classroom assessment practices among the three groups of teachers. The three themes emerging from the data are *Perspectives*, *Policies in school*, and *Professional collaboration and learning*.

Professional perspectives

One factor that helps to explain differences in the nature and quality of classroom assessment among the three groups of teachers is the alignment of teachers' *professional perspectives* to the policy intent. There are two sub-themes associated with this factor. The first sub-theme relates to their views of what their lower secondary students are capable of achieving. The second sub-theme reflects the teachers' views of their roles in teaching and learning, and this is manifested in how they take control and ownership of the curriculum and assessment.

Views of lower secondary students. One influence on the nature and quality of the teachers' classroom assessments is their views about what their lower secondary students are

capable of achieving. The *more aligned* teachers saw their younger students as having the potential to grow and develop. They viewed their students from an asset-rich perspective, and hence created assessments that were scaffolded to provide engaging and challenging tasks. Harry, a *more aligned* teacher, posed higher-order questions to his lower secondary students, even though they did not study geography in primary school. He believed that "it's so much based on the issues you pose to your students." While Totoro was aware that her Secondary 1 students had yet to learn much geography, she continued to present them with challenging tasks. This was because she believed that "the student has that potential and capacity" and "after going through 6 months, 9 months of the learning experience, she will live it up. These girls can excel or do better than this." This belief in the students' abilities to learn and develop was manifested in her school's rationale for allowing lower secondary students to initiate research projects, based on their own interest, under the Student Initiated Assessment (SIA) projects. These tasks involved students proposing a topic of interest, and then venturing out to collect both primary and secondary data to complete the project. These two teachers' beliefs in what their students were capable of are aligned with the types of assessment prompts and tasks they meted out. Both teachers received high AIW scores (Table 6.1), and the work their students completed in response to the tasks also received high AIW scores (Table 5.12).[27]

By comparison, the *moderately* and *less aligned* teachers saw their students from a deficit perspective—they were young and had lacked foundational content knowledge. To this end, the *moderately* and *less aligned* teachers reserved prompts that resonated with the authentic intellectual work criteria for upper secondary students. For instance, Miki would "fall back" to focus on preparing her lower secondary students for tests because she perceived that they were lacking in "maturity" level. Her view was that upper secondary students should be taught with

[27] Only applicable to the first assignment for the work completed by Harry's students.

"a lot more depth" that is "geared to national examinations." This view was reflected in the fact that the classroom assessments she presented to her lower secondary students required mostly short responses. In each assessment, there was just one prompt in the entire task that required students to do some extended writing. Even if higher-order tasks were set, Miki was of the view that "very few groups could do it well." Likewise, Amanda, another *moderately aligned* teacher, felt that setting higher-order questions "depends on whether there is a need at the level to have the higher-order first, because it depends on the student level." In her view, for the lower secondary students "content retention is even not there" and therefore for these students, "you'll have to make it pretty simple." As a result, Amanda was content to use textbook exercises with the lower secondary students because they were "weak students." However, for the upper secondary students, she would set "all those funny questions." One reason was to engage students. Another reason was because the "Cambridge setters also try to be funny." As a result, while more challenging prompts were used, they were to ensure that students were not confused and surprised by the GCE "O" level examinations.

Similarly, Maryanne, a *less aligned* teacher, said that she would mostly provide "challenging questions" for her upper secondary students. This would enable her to "tap their potential and widen the scope for them." However, lower secondary students were "like vessels which I have to pour water into." In particular, she felt her students were "not linguistically competent, and so they might not be able to express themselves well enough." Given such foci, it is not surprising that the *less aligned* teachers designed assessments that did not adhere closely to cognitive demands stipulated in the Assessment Objectives, even though they were aware of the guidelines. For Maryanne, the "ratio is 25 to 75 percent for higher-order to factual questions." She merely followed the required assessment format (i.e., the number of data questions, the

number of short response questions). These views of her students' abilities were reflected in the prompts she used in her assessments, namely fill-in-the-blank items, and in how she required students to read a paragraph and respond to prompts with short answers. For example, the following prompts she used in her second assessment required little inference or analysis:

Figure 6.6
Assessment prompts requiring little inference or analysis

Maryanne, Assessment 2

Section A

- Question 2. Vulcancity refers to _____.

Section B

- Question 3. What is the landform created by this type of plate movement?

Finally, because the lower secondary students might not be taking geography at the upper secondary level, assessments at this initial stage only needed to "expose" students to the subject (James). Typically, teachers like James were not keen to have "any students who are weak in geography taking up the subject" at upper secondary. Aligned with this view, James' classroom assessments were either based on the format that was to be used for the examinations, or he reproduced worksheets from the workbook, so that he could start preparing students early for upper secondary.

Teachers' views of their roles in teaching and learning. The second aspect of teachers' *professional perspectives* concerns their views of their roles in the classroom. One difference in teachers' classroom practices was due to how the teachers constructed their roles in a TSLN classroom. Some teachers saw themselves as being dispensers and controllers of teaching, while others adopted more flexible roles. *More aligned* teachers like Totoro wanted students to deal

with “ambiguities,” and be “comfortable with unfamiliar environments” since the 21st century will be “fluid,” and “answers are not very clear.” In Totoro’s view, correspondingly, teachers themselves also needed to embrace a similar mindset, in relation to teaching and assessment. In terms of assessments like the fieldwork task, Totoro said that “the teacher no longer has control over it.” While teachers could develop skills and competencies during professional development, Totoro said that it was more important for them to “be willing to commit the time and be willing to make mistakes.” This is especially the case because assessing through fieldwork is “something new that [they] are doing.” With 13 classes visiting the farm over a one-week period, students being led by different guides and engaging in a variety of different conversations with the farm employees, Totoro and her colleagues felt they needed to be “fluid” in their marking of the assessments. This was a somewhat new practice compared to looking out for pre-stipulated responses on the mark scheme. The scoring of the fieldwork assignments had to accommodate the range of information that the students gleaned during the trip. As a result, “there are many answers [Totoro] can accept,” based on students’ experiences at the farm. Hence “as long as it’s a valid response, [she] would reward them, and especially when it is context specific.” The variety of responses and experiences also meant that her students had to “be prepared that sometimes answers can be different.” Likewise, Harry would accept a wide range of answers, and his mark scheme took into account “multiple perspectives.” Therefore, when implementing assessment tasks such as independent research and fieldwork projects, teachers had to help students understand that a variety of responses would be accepted, and students needed to realize that they could be given credit for different responses because the marking scheme was broader. Therefore, these two teachers guided their students to understand that in open-ended tasks, there was no one right or wrong response. They also enabled their students to comprehend that they

could not simply rely on the textbook for an answer. By making their students aware that such assessment prompts and tasks can have more than one response, the *more aligned* teachers were guiding students to challenge their views about learning and assessment. The perspectives articulated by these two *more aligned* teachers suggest that they were more attuned to the demands of 21st century learning than their colleagues, and hence, designed and marked assessments that allowed for more diverse responses.

Among the other teachers, only Margaret, a *moderately aligned* teacher, provided for more flexibility in teaching and learning. For instance, she did not want her students to be too worried about right or wrong answers. This is why she departed from the traditional task prompt of requiring her students to define a concept in words. Rather, she asked them to "draw what communications **means to you**" (emphasis as in the teacher's worksheet). What mattered for her was that students tried and could demonstrate what they knew. She wanted to let them know that "marks are not important." Instead, it was the close interaction between teaching and assessment that provided the context for more flexible tasks, and for tasks that prepared students for the *test of life*. This was evident in the prompt in her third assessment, "To ensure a sustainable water supply, Singapore has put in place a system called the four national taps. (a) Name the taps. (b) Which tap do you think is most sustainable and why?" Here, Margaret's mark scheme allowed students to name any of the four national taps. "Once they could justify their opinion, it's fine. I'd give them marks for it," she reflected.

Some of the *moderately aligned* and the *less aligned* teachers saw themselves as controllers of learning. *Less aligned* teachers like James preferred his students to reproduce formulaic model answers. Miki, a *moderately aligned* teacher, also followed her mark scheme closely when grading her students' work. These teachers acted as examiners. As a result, they

were always preoccupied with test construction procedures and processes germane to testing such as "mark schemes," and "types of questions" like "multiple-choice" (Margaret). This also explains why, unlike their *more aligned* colleagues, after these teachers had finished marking, they first asked their students to "check students' marks" rather than to analyze students' responses for misconceptions. Wearing the examiner's hat prompts and spurs the teachers to focus on examination preparation through drumming in correct responses, doing multiple preparation examination papers, and stressing response and test-taking strategies.

Because the *moderately* and *less aligned* teachers saw themselves as examiners, they drilled their students on test taking strategies to accumulate marks (e.g., Jiajia), adhered closely to examination procedures (e.g., Amanda), and focused on examination preparation (e.g., James, Maryanne, Jiajia). For instance, *moderately aligned* teacher Jiajia ensured that her students were "always in touch with trying to extract information" so that they would be able to respond to these questions quickly. Likewise, James and Maryanne were relentless in exposing their students to multiple permutations of the same examination questions, so that students were able to "vomit" responses during the examinations (James). To ensure that his students had facts at their fingertips, James played a game with them at the end of his lessons to ensure that they were able to "remember the textbook." He described how

> at the end of the lesson, if I have time, I get them to name countries, continents, or some random geographical concept that I need them to memorize or give me choral answer. Those who have answered get to sit down. You can see that those who are standing up, they feel the pressure. They will be flipping through the textbook, trying to look out, trying to spot my next probable question and the answer.

James did not see his role as being one of supporting or enhancing learning, but regarded himself as a supervisor who monitored whether students had learned their facts. Therefore, it was not surprising that his tests and classroom assessments included the testing of definitions.

Furthermore, after marking his students' work, his formative assessment strategies consisted of efforts to make his students copy model answers, even if they had obtained the full score.

The teachers' views of their roles in the classroom influenced the way they enacted their curriculum and assessment. As the *more aligned* teachers embraced more flexible practices, they tailored the curriculum to their learning goals, and accordingly, created their own assessments to align with these learning goals. Although there was a national geography curriculum at the lower secondary level, the *more aligned* teachers did not merely adopt the program of study in the sequence presented in the official documents and textbooks, but rather took control of the syllabus, and modified and adapted the program so that learning could be meaningful and purposeful for their students. Consequently, their assessments were aligned with the school-based curriculum that they assigned and matched to their learning goals of developing higher-order skills. For instance, Harry explained that such classroom curriculum implementation was possible at the lower secondary level because teachers had the autonomy to make curricular decisions. This meant that he would sometimes decide not to cover topics listed in the scheme of work. Specifically, in "Sec 2, Sec 1, you can deviate and you do it for a purpose. You want to make geography rich," and teachers have been "given the liberty to relook at how we want to do our testing." Likewise, Totoro adapted the national curriculum to her school's context. Instead of teaching physical geography at Secondary 1 and human geography at Secondary 2 as indicated in the syllabus, Totoro and her colleagues flipped the curriculum to accommodate students' readiness level. The department started this "experiment a few years ago in 2007" because Singapore's urban-living students were finding physical geography "too difficult" and they "were not ready." As a result, Totoro's school made the adjustment and they put "physical

geography at the Sec 2 level and [the teachers] found that they are coping much better." Consequently, the assessments these two teachers designed mirrored their curricular goals.

Allied to their attempts to personalize the curriculum, and in contrast to their *less aligned* colleagues, the *more aligned* teachers did not adhere religiously to the suggested assessment format in the syllabus. Harry decided to be creative and purposeful so that behind his assessments, his students could see the "relevance of all these issues" discussed in the geography class. As I discussed under the section on the teachers' learning goals, the *more aligned* teachers departed from the traditional assessment because they wanted their students to demonstrate their learning from experience or from independent research. Thus, rather than teach particular content areas, these teachers sometimes created assignments that required students to conduct research into them. For two assessments, for example, instead of relying on facts and perspectives from the textbook, Harry created tasks that required his students to conduct research by gathering current views on the issues affecting Singapore. Similarly, instead of conducting a lecture on high technology farming processes, Totoro enabled her students to experience, touch, and observe the processes on the farm, and then to organize this information in a narrative that they constructed themselves. These teachers held the view that it is not necessary to teach everything before assigning an assessment task. For example, Totoro had yet to teach the processes of high-technology farming at the time of the fieldtrip. She tasked her students to do reading and research on the topic, provided scaffolding so that the "frightening experience does not overcome the fun experience."

In comparison, the *moderately* and *less aligned* teachers did not adapt or shape curriculum and assessment to make learning more meaningful for their students. Rather than focus on the learning, the *moderately* and *less aligned* teachers focused on "marks," especially

marks that could be obtained at the end of year or end of key stage examinations. As a result, the *moderately* and *less aligned* teachers taught the curriculum that dovetailed with the upper secondary examinations, and accordingly, the assessments mimicked the format and structure of the standardized assessments. These assessment practices were adopted, even though there was more autonomy and room for classroom-based curriculum development, and despite the fact that the stakes were not high at the lower secondary level. For instance, in their review of the curriculum, Jiajia and her colleagues re-shaped the lower secondary assessment so that the topics taught there would be similar to those at the upper secondary level. In addition to curriculum alignment, Jiajia and her colleagues developed a "big picture" approach during the professional learning team meetings.[28] This approach consolidated the upper and lower secondary curricular and assessment demands into a "progressive framework." This effort was "exam driven" because the framework guided teachers' selection of topics that would be aligned with what was needed in the upper secondary syllabus. For example,

> the emphasis was on the topic of *Population*, because this Sec 2 batch will be sitting for the new syllabus. Hence, I think the topic of *Population* will give them a better foundation, when they go to upper sec. So, we will spend more time on that. In terms of assessment, that is the focus (Jiajia).

Because teaching and assessing at the lower secondary level were planned "according to the expectations of the upper sec" (Amanda), students taught by the *moderately* and *less aligned* teachers experienced a narrowing of the curriculum. In comparison, students taught by the *more aligned* teachers were exposed to rich geographical experiences through the curriculum and assessment. Furthermore, the students taught by the *more aligned* teachers had the opportunity to be assessed in a variety of ways through different modes and assessment formats, whilst their

[28] According to Jiajia, in Singapore, schools are a professional learning community. The groups in each school form professional learning teams.

peers were engaged in assessment activities that mirrored the format of the end of year or high stakes examinations.

In summary, the differences in classroom assessment practices among the three groups could be partly explained by the teachers' perceptions of their students' capabilities and abilities, and by the way they conceived of themselves as teachers. While TSLN and TLLM envisioned that all students can achieve different "peaks of excellence" (Shanmugaratnam, 2007), the majority of the teachers in the study continued to embrace "hereditarian theories of intelligence" (Shepard, 2000, p. 7) and held limited views of student development in which only upper secondary students could be engaged and be given challenging tasks. This resulted in the *moderately* and *less aligned* teachers having lower expectations of their lower secondary students, and setting assessments that did not require students to demonstrate higher-order skills.

Policies in school

The size of Singapore's education system and the small geographical space of the country suggests that it is easy to hold face-to-face communication between the Ministry of Education and school leaders. In fact, Singapore's 360 school principals are able to converge within one hour's driving time to meet with policy makers when new initiatives are announced (A. Hargreaves & Shirley, 2012). However, the interview comments suggest that, after the macro intent has been communicated, meso or school level policies mediate and influence the fidelity of implementation in the classroom.

TSLN recognizes that education in Singapore can no longer adopt a one-size-fits all model. Rather, schools have been conceived to be "crucibles" of ideas and innovations (C. T. Goh, 1997) where teaching and learning are tailored to the profiles of students in each school. At the macro level, the national curriculum was designed to reflect TSLN's goal of developing

higher-order skills, and the syllabuses used assessment objectives to signal a balance between knowledge recall and higher-order skills, and to encourage teachers to use a variety of assessment types. At the *meso* level, schools were places of curriculum innovation where teachers could use "white space" to customize the curriculum (Shanmugaratnam, 2005a). At the *micro* or classroom level, the *more aligned* teachers drew on the macro policy direction and adapted this to their school context. As the interview data and AIW scores suggest, some teachers, more than others, enacted classroom assessment practices that resonated more with the TSLN vision and the TLLM tenets.

However, the translation of the policy into the classroom is more complex. The *meso* layer, or the school level, is an additional layer that facilitates or hinders the policy realization in the classroom. At this middle level, schools need to provide support and resourcing for teachers. School policies can directly or indirectly influence and impact teachers' assessment practices.

One school level policy that directly influenced teachers' assessment practices was teacher evaluation. While TSLN and TLLM focus on developing students for the *test of life*, some schools appeared to continue to require that their teachers prepare students for a *life of tests* at the time of this study—fifteen years after TSLN was initiated. The teacher accountability and evaluation approaches adopted in some schools resulted in teachers such as James and Maryanne (*less aligned* teachers) continuing to use classroom assessments that detracted from the TLLM tenets and TSLN vision.

Teachers from the three groups spoke of, and were cognizant of a teacher evaluation instrument their schools employed to examine if students had benefitted from attending the school's academic program. James explained that this was

> the Mean Subject Grade. When they enter secondary school, [students] have this MSG following them around, based on their PSLE[29] scores. For Express classes, their MSG for geog is around B3, 3.5. So all of them are expected to get B3 and above. This pressurizes the teacher. Because if their grades drop below the MSG, it means that I'm not value-adding, and I'm not delivering.[30]

A good MSG score must be low and Maryanne's reporting officer constantly reminded her that "the lowest possible is better." Therefore, for James, "at the end of the day, I think I'm pressured to deliver results, to test them on what is in the textbook, syllabus, and all that stuff." He observed that "it's hard for me to find a middle ground between making my lessons interesting, or just trying to drive my points across so that when they sit for exams, they can score." Although this measure of student progress "pressurizes the teacher," James saw its value and purpose: for him and for his school, it meant that they "want to value-add [their] students." Miki, a *moderately aligned* teacher, explained that students who enter her school with high PSLE scores but do not do well, teachers "need to find out why."

Some schools' high-stakes use of this data results in intense pressure on teachers. For Maryanne, the pressure was greater because the MSG computation was part of her teaching "portfolio" and her reporting officer would use it for the "ranking." At Maryanne's school, "they actually look at it across the board." While schools might use the MSG measure to rank and rate teachers' performance, it typically applied to teachers whose students were sitting for the GCE O-level and N-level assessments. The practice in Maryanne's school appears somewhat harsh because the MSG is computed and measured, "even for lower sec,"—a non-key stage level. James, another *less aligned* teacher, experienced the same pressure. He said that

---

[29] The Primary School Leaving Examination (PSLE) is an assessment all students take at the end of primary education. The aim of the PSLE is to place students in different secondary school courses (Lim & Tan, 1999).

[30] Adapted from personal correspondence with Rita, the senior teacher: In the Singapore examination parlance, a passing mark is 50 (or grade C6), as typically all assessments are based on 100%. The top grades are A1 and A2 and students need a score of 70 to obtain an A2. An A1 is worth about 75 marks. A B3 grade constitutes a score of about 65.

> The least I can do is to meet the MSG. It means they are at that standard that they came in. But if I can value-add on the MSG, instead of 3.5, they get a 3.1, it means I am doing better.

Given schools' use of this instrument, it is therefore not surprising that some or all of the assessments that the teachers participating in this study submitted mirrored the examination format suggested in the syllabus document. This was true even for the *more aligned* teachers. Harry and Totoro's common tests also resembled the recommended test blue print. Perhaps because their schools were more stringent and explicit in the use of the MSG as an indicator of teacher effectiveness and the quality of student learning, the *less aligned* teachers felt more pressure to churn out results, and this, in turn, affected the types of assessments that they used. Compared to the *more* and *moderately aligned* teachers, the three assessments James and Maryanne submitted did not deviate from the format mentioned in the syllabus. The worksheets they used and described were taken from the textbooks and workbooks. (See excerpts from James' assessment in Figure 6.5 and examples of Maryanne's assessments in Figure 6.6). To indicate to their supervisors that they were working toward raising or maintaining students' MSG, Maryanne and James did not stray from what was prescribed and decided on by the school and their departments. Instead, they strictly adhered to the test blueprint to signal to their reporting officers that they were not deviating from that which was required of them. As presented earlier, James struggled with finding time for formative assessment; he did not have time to re-teach or allow his students a second time to rework the tasks because he had to press on with his department's scheme of work. Ironically, this work plan was created to be efficient, so that if a teacher from the school were to be absent, the "relief teacher will just need to pass the piece of work to them and get them to do it." However, this structure, intended to ensure that teaching

progressed efficiently, left teachers with insufficient time to provide students with support and remediation.

In comparison, the other teachers had assessment practices that reflected TSLN's intent more closely. As mentioned earlier, the *more* and *moderately* aligned teachers used a wider variety of tasks than did the *less aligned* teachers. Furthermore, as reflected in the AIW scores, the *more aligned* teachers assessed a larger range of skills. One reason for this was that in comparison to the two *less aligned* teachers, the other teachers did not work in schools that used high-stakes assessment data at the lower secondary level. Like James and Maryanne, Amanda—a *moderately aligned* teacher—was cognizant that her school administrators monitored student progress using the MSG measure. However, she did not feel the intense pressure that Maryanne and James described. Perhaps this is because Amanda's school's administrators were only "suspicious when students perform beyond the normal range." She described two instances in which the teachers were questioned.

> Maybe only 60 percent passed. This is not enough for the Sec 1. And they will ask if the paper was difficult. So we have to do the markers' report (Amanda on an instance when students did badly).
>
> There was one year when we did very well—Wow! The MSG for the Sec 1 was like 1 point something—then they also questioned. They said, too simple, your exam paper. Only the extreme, then they really follow up (Amanda on an instance when students performed extraordinarily well).

Similarly, Harry—a *more aligned* teacher—was also aware of the use of MSG in his school. However, like Amanda, he was unperturbed and his assessment and learning goals focused on getting his students to appreciate geography as a discipline, and to enable his students to develop higher-order skills. Perhaps this was because his school did not exact the same level of pressure on him to perform as James' school did. Indeed, Harry's school supported his efforts to develop inter-disciplinary efforts with a colleague. Both Harry's and Amanda's experiences

indicate that there were a variety of accountability practices and consequences across the schools with regard to the use of the MSG. Nevertheless, it is clear that the penalties attached to this MSG value-added measure made the teachers anxious. James and Maryanne explicitly highlighted these pressures and stresses. For instance, James mentioned he was pressured to deliver results, and Maryanne's department head frequently reminded her to ensure that she obtained a low MSG – an indication of good student scores – for each examination.

Thus, aligned with TSLN's vision of schools being "crucibles" of ideas and innovation where teachers and principals "constantly look out for new ideas and practices, and continuously refresh their own knowledge" (C. T. Goh, 1997), teachers who worked in schools which trusted their staff to go about their professional duties (e.g., in Amanda's and Harry's schools) and which broadened teacher evaluation criteria had assessment practices that reflected the policy intent more closely. By contrast, teachers who worked in schools that continued using student performance as an indicator of teacher effectiveness were constrained in the way in which they approached assessment and learning.

Another aspect of school policy that influenced classroom assessment is administrative support for teachers' efforts in shifting and changing their classroom assessment. Where there was greater administrative support, the teachers could design and implement more innovative alternative assessments. Administrative support could be in the form of time tabling or assigning staff to support the conduct of fieldwork assessment. For example, Totoro's reconnaissance trip helped her to make curricular, assessment, and logistical plans. Her school also provided support by releasing other staff from their duties to accompany Totoro and her students on the trip. Similar administrators created time for Harry and his colleague to meet and plan their integrated task, and this support was provided consecutively for three years as the assessment Harry

contributed was into its third iteration. Assessments that require the integration of knowledge across two or more disciplines require teachers from different departments to meet, discuss, and collaborate. Harry's research assessment was co-created with a colleague and this required time for discussion and for making modifications. After several attempts, the teachers achieved the high quality task that was submitted for this study. In these two examples, Totoro and Harry came from schools that encouraged collaboration across departments rather than "balkanization" (A. Hargreaves, 1994) of the individual subject departments. Consequently, Harry could work with a colleague from the history department to design a research assignment, while Totoro could rely on colleagues from her department and from other departments to help her lead the fieldtrip.

In comparison, the schools in which the *more* and *moderately aligned* teachers taught did not provide the same level of support for alternative assessment. Because of the nature of school support, Miki and Maryanne lamented the lack of "time" to engage their students in authentic intellectual work. Miki was unable to set more alternative assessments for her students because "it takes a lot of time." Maryanne was also reluctant to get her students to engage in "oral presentations" because "they take up too much time, which we don't have." And unlike Totoro who was able to garner colleagues' support to accompany her thirteen classes of students to the farm on different days, Maryanne did not frequently use fieldwork to assess her students because of "manpower constraints." Teachers like James and Amanda did not even discuss alternative assessments. To this end, even though some of the *moderately* and *less aligned* teachers wanted the opportunity to present their students with authentic intellectual work, they were unable to do so because of the lack of school support.

School support was necessary to provide sufficient technical support for the teachers to conduct alternative assessments. For instance, Harry's independent research task required students to submit their responses online via an IT platform and in a hard copy. This meant that teachers needed contingency plans for bottlenecks such as students not being able to access the online portal, and ensuring that the printers in the school were ready to make copies. In addition, Harry said that they had to bear in mind that students from low social-economic backgrounds might not have easy and ready access to a computer and thus the teachers would need to make provisions for them. For Totoro, planning for the trip to the farm involved a huge logistical exercise as 13 classes, each with about thirty students, were scheduled to visit the farm. The logistical planning for a fieldtrip also included seeking colleagues' assistance in accompanying students to the farm, sorting out the budget, and making the appropriate arrangements with the farm. To this end, teachers require substantial amounts of planning time and logistical support to implement these research and experiential assessments.

School policies are typically introduced to increase efficiency, manage teachers' workload, and raise the rigor in teaching and learning. However, these policies inadvertently affect the nature and quality of teachers' assessment practices. All the teachers—including the *more aligned* teachers—discussed how they had large cohorts of lower secondary students. In some schools, one teacher was responsible for the entire cohort (e.g., for James). In other schools, there were several teachers teaching the cohort (e.g., in Amanda and Margaret's schools). Because of the cohort size, teachers across the three groups shared a common viewpoint, that is, that the tests need to be fair and objective for all students. The tests had to be pitched to the middle ground so that the weaker students "are not left behind" (Amanda).

To ensure fairness and efficiency during testing and assessment, more and more schools were centralizing and standardizing the teaching plans and assessments to manage teachers' workloads, and to ensure comparability across classes. "In the past, when there was not so much coordination, whatever we want to set for the class, we set" (Amanda). However, because of the variations in practices, expectations, and standards, schools started to coordinate assessments. Amanda explained that "we need to be a little more aligned."

Amanda's school saw the need to standardize the assessments to be fair to all students and to ensure rigor. The intent of these new procedures was increased fairness and fewer variations in the quality of the assessments. There were advantages in standardization because all students in the level could attempt the same assessment. On the other hand, teachers teaching in schools such as Amanda's were now compelled to rush through the syllabus to ensure that their classes were up to speed before the next mandated end-of-topic assessment or common test. This suggests that teachers like Amanda had less autonomy than colleagues like Harry to shape instruction to meet their students' needs. This is why, as James explained, there was little time for him to re-teach or to address students' misconceptions. Because of the planned scheme of work, he had to "proceed with the set assignments, even though [some students] might not be ready for it."

These arrangements had implications for assessment. These school-based assessments became mini-standardized tests, developed under secrecy and implemented *en masse* to students at appointed times. This indicates that assessment practices were less customized but rather pitched at a generic level. This was one of the reasons why teachers questioned how they would stretch their high-ability students without leaving the low progress learners behind. For instance, after each assessment, Totoro reflected that "I may not be able to reach out to each and every kid"

but as long as "they don't feel that I insult their intelligence," the assessment can be deemed suitable. Amanda faced a similar constraint too as she felt that she was "not stretching [her] best class." As a result, she was open to constructing more creative questions for the lower secondary students since at this level, the teachers could afford to "be a bit more adventurous." Ironically, administrative changes to lessen teachers' workload (e.g., in James' school, a coordinated department plan to ensure that materials and tasks are ready for teachers covering lessons) and to ensure more equitable assessments standards by standardizing tests in schools, appear to have contributed to reduced customization of teaching and assessment in relation to students' needs, thereby resulting in practices being less aligned to the TSLN intent. This practice of standardizing tests in schools might help explain the large number of common tests submitted for this study.[31]

In summary, the analyses pertaining to *Policies in school* suggest that structures and the contexts in which teachers work do not necessarily support or encourage teachers to adopt change and innovation. First, teachers who worked in schools that did not embrace TSLN's focus on preparing students for the *test of life* continued emphasizing examination preparation and grades. This was because their schools continued to evaluate their effectiveness based on the types of grades their students attained. Second, school policies that were designed to lessen teachers' workload and to present students with assessments of comparable difficulty and quality, ironically, resulted in standardized assessments that resembled miniature high-stakes assessments. This led to teachers using assessments that were less customized to students' needs and ability levels. Consequently, more than a decade after TSLN's implementation, its aspirations still are not evident in many classroom assessment practices.

[31] As reported in Chapter 5, half of the 24 assessments submitted to this study were common tests, class tests, or examinations.

Professional learning and collaboration

The third theme that provides explanation for the differences in classroom assessment practices among the three groups of teachers is *professional learning and collaboration*. Professional learning opportunities enable teachers to change their practices through leveling up competencies in and exposing them to new ideas and practices in the areas of curriculum design, classroom instruction, and assessment (Wong, 2007). There are benefits when teachers attend professional learning sessions to use formative assessment strategies (e.g., Black, et al., 2003b; Dixon & Haigh, 2009) or to design assessments that encapsulate the AIW criteria (e.g., Avery, et al., 2001; Koh, 2011b; Koh, et al., 2012). For example, participating in a project on formative assessment strategies enables teachers to a develop deeper understanding of the nature and purpose of feedback, as well as to re-examine their roles in the classroom (Dixon & Haigh, 2009). Furthermore, sustained professional development to guide teachers in designing authentic classroom assessments and rubrics improves teachers' assessment literacy (Avery, et al., 2001; Koh, 2011b; Koh, et al., 2012). Through participating in professional learning, teachers develop a common language to discuss their assessment practices (Avery, et al., 2001). At the same time, while there are changes in teachers' attitudes following professional development sessions, more time is needed for teaching practices to change (Dekker & Feijs, 2005), indicating that professional learning needs to be conducted over a period of time rather than through a once-off session (Avery, et al., 2001; Koh, et al., 2012).

The teachers participating in this study also indicated that professional learning in the area of assessment is instrumental in the kinds for assessment they create. Through professional learning sessions, teachers were introduced to new skills and ideas. These sessions also provided opportunities for teachers to address and clarify questions relating to the construction of tests and

other administrative procedures. The latter purpose of professional learning was particularly true for the *more* and *moderately aligned* teachers.

Professional learning provided the *more* and *moderately aligned* teachers with theoretical knowledge that guided them in developing tasks that embodied AIW-type criteria to reflect TSLN's intent. These teachers also diligently applied the skills and ideas they picked up to change their assessment practices. From their in-service learning sessions, Totoro and Jiajia did not see summative and formative assessment as diametrically opposed. Totoro saw the two concepts as supporting each other, and Jiajia sketched a diagram to show how these were part of teaching and learning (Figure 6.2)**.** This indicates that some of the professional learning sessions were designed to reflect the TLLM tenets on assessment. In addition, the *more* and *moderately aligned* teachers valued formal and informal professional learning. They considered school-based or MOE-organized sessions as formal learning, while informal sessions included discussions with colleagues, especially with more senior teachers. For a few teachers, professional learning even included participating in this research study. For Totoro, talking about and reflecting on her assessment practices during the study was a novel experience, and she wondered why "I have never really talked about my assessment thinking aloud like this. And in such a detailed manner." In terms of formal professional learning, Totoro who received high AIW scores for two of her assessments had the most consistent and intensive training. She attributed her deep knowledge to "internal and external" training conducted by MOE and sessions organized by her school. The fact that her "school has trained us well" was evident in the way she discussed her views about assessment as well as in the intense and thoughtful way she analyzed the assessments she contributed to the study. Likewise, Harry saw professional learning as being the means to help him "become a better teacher" and, indeed, as the "overall

in-charge of the [school's] key training programs," he "encouraged" colleagues to actively learn from one another or from formal professional development platforms.

For *moderately aligned* teacher Jiajia, professional learning is the "critical enabler" in building her assessment literacy and enhancing her formative assessment practices. To this end, in her school, everyone "whichever course or workshop that they have attended, they just come back and share." Margaret, another *moderately aligned* teacher, found that she could pick up powerful techniques that enabled her to focus on learning rather than on testing. As a result, she was consistent in and committed to applying useful ideas she gleaned at the workshops and conferences she attended. For example, from a conference she attended, Margaret picked up the idea of constructing assessment prompts that did not indicate the marks to be attained. The speaker had suggested that when marks are not attached, teachers can get students to "go a little deeper" in their responses. This is because when marks are used, "all students look for is the mark when teachers give back the test." With this in mind, Margaret decided to move from the conventional practice of using marks towards a comment-only assessment for her second assessment on "Communication," in which she asked students to define this concept in text format and as a graphic. Because some concepts are very abstract for students, and because "some students are actually better at drawing pictures," she decided to apply this assessment prompt to provide an alternative way for students to demonstrate their learning. When reflecting on her students' responses to the assignment, Margaret observed that

> They kind of felt like they had more ownership of their own learning—because I was just asking them to tell me what you know. As I told them before, I hadn't taught this yet, so nothing's right or wrong. You just have to let me know what you know.

It was reflections such as these that led to Margaret being placed in the *moderately aligned* category, despite the fact that her AIW scores were the second lowest among the eight teachers.

However, the *less aligned* teachers saw professional development sessions as avenues to find out the best way to answer examination questions. James expressed his disappointment at one assessment workshop organized by the examinations office because the trainer had not provided clear answers on how students were expected to answer questions

> There will be Q&A sessions at the end of the day, and they are asked about the level of response questions and all that stuff. But their answers are so generic! "Do you need a conclusion?" and they'll say, "No, you don't need a conclusion." But then, the question asks, "Do you agree?" So shouldn't you give marks for the conclusion? They don't really answer the question, the guys from [the curriculum office]. They say, "No, as long as [students] compare and evaluate, they give the advantages, the limitations, they give examples, you should give them maybe as high as a level 3 mark—7 or 8 marks." Then what about the "Do you agree part?"

As compared to the *more* and *moderately aligned* teachers who viewed professional learning as opportunities to learn more about different assessment techniques or to expand their repertoire of assessment skills, James wanted "assessment courses to give us samples that would make it easier for us to craft our assessments, and work backwards from here—our lessons, our assignments." He was annoyed when "the [examination board] chooses something brand new that our students are caught off guard." James' view of professional learning mirrored his learning goals—to find ways and means to obtain model examination answers and prompts, rather than to focus on developing learning. Therefore, the critical element is not just attending professional learning sessions, but rather the objective of attending professional development, and subsequently how the teacher applies the learning in class. The latter is what ultimately determines the nature of classroom assessment practices that are enacted.

Another aspect of professional learning that enhanced and supported teachers' classroom practices was peer collaboration. Learning from peers was useful in developing teachers' skills in constructing assessment tasks. When explaining why he picked the assessment, Challenges Singapore Faces, as the piece that most exemplifies the TLLM tenets, Harry enthusiastically

discussed the value of working closely with a colleague to design and refine the task over a period of three years.

> That's why I really love this piece of work, especially when you work with another colleague. My colleague gave a lot of feedback. Feedback is given on Thursday, and we take home over the weekend, edit and do a lot of things. Then we asked ourselves, 'What's the best way?' So it's very much scaffolding. The words are big, what they are supposed to do. And even this one here about the challenges Singapore faces, she gave her point of view in terms of a history teacher. I gave my point of view as a geography teacher. I think we could see the whole, big, big, big picture.

As a result of this close collaboration, the independent research assignment submitted for this study was, and is one of which Harry was exceedingly proud of, and is into its third iteration. Through iterative revisions and peer critique, it was not surprising that this task received the highest AIW score (Table 5.8). Harry said that this was because "we sat down and we analyzed and we thought about it." This collaboration was powerful, even if it took "15 to 20 hours" for the third iteration of the task. The value of peer learning was why novice teachers like Margaret would like it if "a more experienced teacher, when I come up with an assessment, kind of sat with me and went through it together." Such "peer learning" would be "really effective" and hence "a lot of the time, [she'd] go to other colleagues, ask their opinion on the assessment." In her view, the benefit was that "everyone has different ideas. We can learn from each other."

Peer collaboration was powerful because it involved colleagues asking one another critical questions to challenge thinking and to adopt new perspectives, as reflected in Harry's comment above. Peer collaboration involves teachers learning from each other. It amounts to more than teachers working together to ensure administratively that appropriate test construction protocols are adhered to, which are processes adopted by Amanda, Maryanne and Miki. Amanda vetted assessment prompts designed by younger colleagues. Miki would confer with a college to decide if prompts were manageable for the lower-ability students. Maryanne explained

that for test development, she and her colleague "will confer" and discuss personally or "ding dong each other through email" to decide on the topics and format. The experiences mentioned by Amanda, Maryanne and Miki were different from the critical friend support that Harry and his colleague provided for each other. In the case of these other three teachers, the discussions were purely administrative and procedural. Conversely, for Harry and his colleague, the experience from the collaboration was mutually beneficial to their professional learning.

The *more aligned* teachers and some *moderately aligned* ones valued personal self-reflection as a means to improve their assessment skills, especially in assessing higher-order skills. Miki and Jiajia reported that the in-depth reflections on their assessments at each interview enabled them to develop new insights into their practices in geography assessment. Participating in the study benefitted them as they engaged in intense reflections and discussions to critique and review their assessments. At the end of the six-month study period, some teachers said that they now spent more time pondering their decisions related to classroom assessment. They also devoted more time to analyzing the assessment information, and designing the assessment prompts. Jiajia observed that the discussions on higher-order skills made her think more deeply about whether she had "been testing it, assessing it." She also questioned whether her prompts were "considered higher-order thinking skills." Miki felt that she became more conscious about her assessment practices over the course of the study. She said that she had been "put … on [her] toes a little, because [she knew she was] going to be interviewed and surveyed" and as a result, would "try [her] very best to design questions and have feedback, have analysis, do more than what [she has] done before." Totoro remarked that the interview sessions had made her "more conscious" about her practices and the types of decisions she makes. In particular, "when I look at the paper, somehow I tend to be more critical." It was through these

reflections that some of the teachers realized that peer critique and reflection of their assessment practices was not an embedded or pervasive aspect of their practice.

While teachers do not generally critically review or discuss their assessment questions with their colleagues (Black & Wiliam, 1998a), the data compellingly illustrate the value of professional learning when teachers are trying creative and innovative programs and practices in schools and classrooms. Learning in a community such as in Harry's and Jiajia's is characteristic of socio-constructivist theories in which situated communal settings enable learning to be distributed within the community (Lave & Wenger, 1991). In this learning process, each teacher's "human capital" becomes a "collective social capital" and the group is "building the capabilities" together (A. Hargreaves & Fullan, 2012). Such learning opportunities enhance teachers' assessment competencies, as evident in the example of Harry and his colleague.

In summary, the differences in the teachers' assessment practices are partly explained by issues of *professional learning and collaboration*. Teachers who participated actively in professional learning, and who engaged colleagues in professional dialogue enacted assessment practices that were *more aligned* to the policy. In particular, the teachers with higher AIW scores or whose comments reflected the TSLN intent were those who used professional learning to deepen their knowledge of a range of assessment theories and practices. These teachers also drew on and applied ideas they acquired at professional learning sessions when developing their assessment tasks. By comparison, teachers who attended professional learning sessions to find out more about test taking and to collect model responses, used assessment practices that were *less aligned* to the policy vision.

Summary

The three themes discussed suggest that the variations in the classroom assessment practices of the three groups of teachers are influenced by these teachers' views about teaching, learning, and assessment, by the school and community they work in and with, and by the professional learning sessions they attend. These themes affect the extent to which the teachers' assessment practices require their students to exhibit skills deemed important and necessary for the 21st century, skills that are aligned to the TSLN vision.

First, teachers' *professional perspectives* about their students and about their roles in teaching and learning influenced the nature and types of classroom assessments they used. The teachers whose assessment practices provided opportunities for students to question and query what they were taught, to gather and interpret data, as well as to critique and analyze facts and information saw themselves as facilitating learning. These teachers also viewed their students as having the ability to learn and hence, provided challenging tasks where possible to stretch their students. Rather than provide a set of model answers to classroom assessment tasks, these two teachers guided their students to figure out mistakes themselves. Comparatively, teachers who saw themselves as examiners focused their attention on test preparation strategies. The *moderately* and *less aligned* teachers who fell into this grouping viewed their lower secondary students from a deficit perspective, they were content with assigning assessments that focused on recall and on definitions, preferring to present higher-order tasks such as analysis and evaluation to upper secondary students. In terms of their formative assessment practices, these teachers focused on dispensing model answers to their students because they perceived their students to be lacking in factual knowledge and skills, and to be requiring a lot of practice tests.

Second, *policies in school* can support or hinder teachers' uses of assessments that focus on higher-order thinking skills. On the one hand, when schools and administrators provide support through resourcing and planning procedures, teachers are able to collaborate to construct and conduct alternative assessments such as fieldwork reports and projects. On the other hand, some school policies such as those on teacher evaluation indirectly influence teachers' assessment practices. Policies relating to teacher evaluation in some schools were accompanied by severe consequences for teachers whose students did not perform at or above expectations. As a result, teachers were less willing to innovate and try a variety of assessment types or to assess their students in areas other than those from the syllabus. For instance, the pressure on teachers to produce results through test preparation was intensified when schools, such as those where Maryanne and James were teaching in, made high-stakes use of the value-added measure, MSG, to evaluate teacher effectiveness. The two teachers responded by exposing their students to multiple permutations of test prompts so that they would be able to answer test questions mechanically or in James' words, "like robots," so that they are able to produce correct responses. In this respect, the *meso* or school level enactment of a *macro* policy tool suggests that there are policy contradictions that result in teachers' assessment practices diverging from the TSLN intent.

The *more* and *moderately aligned* teachers were also aware of the use and impact of this teacher evaluation measure. Yet, in comparison to James and Maryanne, the *more aligned* teachers used a wider variety of assessments types and formats, assessed higher-order skills, and provided chances for students to make sense of their errors. They also did not subject their students to numerous permutations of how a topic or concept could be tested. In fact, neither teacher spoke about retests during their three interview sessions. Likewise, although the

moderately aligned teachers also were also aware of the value-added measure and its importance, they were under less pressure than the *less aligned* teachers because their reporting officers did not relentlessly remind them of the need to ensure good quality grades. As Amanda and Miki mentioned, the administrators in their schools would only require a report if student performance was unusually good or abysmal.

Amidst these cross currents of influence, the *moderately aligned* teachers felt the pressure to ensure their students performed well in tests and examinations because of the large number of tests submitted for this study. In particular, the three assessments Amanda and Miki submitted were common or class tests. One of the three assessments Margaret submitted, for example, was a test for students to practice examination skills. While all teachers spoke about the need to ensure that students perform well, the fact that half the assessments submitted for this study were examinations, and that some teachers only submitted common tests, indicates that there is a continued preoccupation with test preparation and grades among some teachers. On the one hand, the *moderately aligned* teachers such as Jiajia, Miki and Margaret spoke enthusiastically about wanting their students to develop an appreciation of geographical education, and where possible, they wanted to focus on and assess what matters. On the other hand, as a group, the *moderately aligned* teachers ended up focusing intensively on what will be measured. This was seen in the example of Jiajia who worked with her colleagues to judiciously align the lower and upper secondary syllabuses, to the extent that they focused only on teaching and assessing topics that would be assessed in the high-stakes assessment at the next level. Therefore, teachers had to compromise their goals to develop attitudes and dispositions in their students even though they also wanted their students to be familiar with examination type prompts. Efforts to address these different assessment principles at the same time were evident in teachers' practices of including

one or two cursory prompts that addressed higher-order skills or geographical values within the entire assessment task.

Third, *professional learning and collaboration* influenced teachers' assessment practices, and helped them improve the quality of the tasks that they designed. This is a finding that is also reported in the research on the impact of professional development on teachers' formative assessment and authentic assessment practices. For example, in a study of Scotland's *Assessment is for Learning* (Aifl) policy, Hayward and Spencer (2010) reported that teachers valued peer collaboration as a lever that enabled them to re-perceive their formative assessment practices. Likewise, sustained professional development to guide teachers in designing authentic classroom assessments and rubrics improves teachers' assessment literacy (Avery, et al., 2001; Koh, 2011b; Koh, et al., 2012).

The *more* and *moderately aligned* teachers spoke enthusiastically about learning and collaboration. These two groups of teachers used professional learning to acquire new ideas so as to improve their classroom assessment practices, and applied what they learned to the assessments they designed. However, the *less aligned* teachers did not find this aspect helpful for their professional work. By contrast, the *less aligned* teachers merely saw workshops as opportunities to obtain examination tips from central office staff or examiners.

In terms of collaboration, some teachers (e.g., Miki and Harry) spoke of constructing assessments with colleagues. They collaborated with colleagues teaching the same subject (e.g., Miki) or with colleagues from other departments (e.g., Harry). Other teachers (e.g., Jiajia) benefitted from being in learning communities where teachers shared ideas gleaned from professional learning courses with one another. In schools where there was a weak culture of collaboration, novice teachers like Margaret reflected that more collaboration and consultation

would both enable her to design better assessments as well as be beneficial to her professional growth. Comparatively, the other novice teacher, James, did not mention collaboration, perhaps because he was the sole teacher overseeing the entire level. He also did not consult his colleagues because he disagreed with their assessment approaches.

Discussion and Conclusion

The Teach Less Learn More (TLLM) movement was introduced in 2005 to realize the Thinking Schools, Learning Nation (TSLN) vision of developing generations of Singaporeans as thinking and committed citizens able to make sound decisions to ensure that the country continues to be vibrant and successful (C. T. Goh, 1997). The TLLM tenets prompted teachers to revisit *why*, *how*, and *what* they teach. I have suggested that these tenets resonate with features of constructivist learning theory, particularly in the way they call for teachers to do *less* "telling," and *more* "modelling, guiding, and facilitating," to focus *less* on dispensing information but to teach *more* for understanding, and to use *more* "formative and qualitative assessing." According to Biggs (1996a), qualitative assessment, which TLLM envisages, differs from quantitative assessment in that it charts longitudinal growth. Thus, qualitative assessment is developmental in that the learning outcomes are the constructions students make at any given stage. Another feature of qualitative assessment is the use of authentic tasks that require students to work on problems that they would encounter in the real world (Biggs, 1996a). Biggs (1995, p. 4) terms such assessments as "ecological." The characteristics of qualitative assessment resonate with constructivist assessment as well as with Newmann and Associates' (1996) AIW criteria. In comparison, quantitative assessments require students to reproduce previously learned material quickly and correctly (Biggs, 1996a). Based on these descriptions, qualitative assessment is aligned with constructivism while quantitative assessment reflects behaviorism.

Aligned to the research questions, the aim of this chapter is to understand and suggest explanations for the assessments which the teachers submitted for the study, and which were examined based on the AIW criteria in Chapter 5. First, the analyses of the teachers' comments were a means to understand how the teachers elicited student learning. Second, the data were analyzed for the ways teachers engaged in formative assessment in order to comprehend how they enhanced student learning. Third, this chapter sought to elicit possible reasons to understand the types of assessments the teachers used. To this end, this chapter drew on an adapted constructivist assessment framework to analyze the interview comments of the eight teachers participating in this study. The analysis also referenced the TLLM tenets. Based on the analyses, there were three categories of practices—*more*, *moderately* and *less aligned* to TSLN. In general, teachers whose assessment practices resonated most with the policy frequently described and used assessments that reflected the policy intent, and embodied the characteristics of constructivist theory. Conversely, teachers whose practices were less aligned to the policy and constructivist theory mentioned or focused on aspects that were least consistent with the policy intent.

From the data, three possible explanations emerged for the variations across the three groups. First, the differences in the practices among the three groups could be explained by the extent to which there was alignment between teachers' perceptions of their students' abilities and capabilities and of what their roles as teachers were supposed to be within the policy intent. In particular, teachers whose views of their roles and of their students were closely aligned with the policy and its undergirding constructivist principles enacted assessments that most reflected TSLN's intent. Second, school policies that were aligned to TSLN's vision of encouraging change from the ground up, and of reducing the emphasis on testing, enabled teachers to

construct assessment practices that resonated with the policy intent. Third, school-based professional learning and collaboration were platforms and opportunities for teachers to explore the use of assessment formats and modes that departed from those that had been used traditionally.

This section discusses the significance of the findings gleaned from the interpretation and analysis of the teachers' interview comments. Specifically, three themes are discussed: (1) Patterns of assessment, (2) Learning versus achievement, and (3) Understanding teachers' assessment practices.

Patterns of assessment

The teachers' comments support and lend explanation to the micro level data analyzed and presented in Chapter 5. From the 24 assessments the teachers submitted, three patterns of assessments were reported in Chapter 5. First, the eight teachers used a *variety* of assessments in different formats and types, comprising in-class assessments, examinations, class tests, independent research and a fieldwork task. Second, there was *persistence* of practices like the testing of facts and knowledge, despite the fact that TSLN was launched as part of MOE's recognition that in the 21st century, the goal of education was to equip students with skills and habits of learning that will facilitate lifelong learning (C. T. Goh, 1997). The teachers' persistence in assessing discrete factual knowledge was indicated by the higher scores for the *Disciplined Inquiry* criterion and lower *Construction of Knowledge* criterion. As mentioned in Chapter 4, high *Disciplined Inquiry* scores indicate that teachers focused on content knowledge instead of disciplinary skills and thinking, while low *Construction of Knowledge* scores meant that teachers were not requiring students to make sense of and interpret data pertaining to an issue. Third, the *quality* of the assessments (or the extent to which the assessments assessed

higher-order skills) was low as indicated by the AIW ratings, in which just 7 of the 24 assessments had scores above the midpoint of the scale. The low AIW scores suggested that teachers did not require their students to demonstrate higher-order skills, such as those represented by the AIW criteria. Instead, the scores suggested that the teachers emphasized the testing of disciplinary or content knowledge. The interview data presented in Chapter 6 provide further insight into and understanding of the nature, purpose, and types of assessments contributed by the teachers in response to TSLN. Similar to the assessment patterns presented in Chapter 5, the teachers' interview comments were grouped into the three categories of *variety*, *change*, *and persistence*.

Variety. The teachers' comments suggest that teachers used a *variety* of assessment practices, ranging from strategies that were aligned to the policy at one end to approaches that detracted from the reform on the other. In addition, some teachers used more assessment types and formats than others. Specifically, the *more aligned* teachers used the largest range of assessment types, as they spoke about using formal assessments like common tests as well as informal assessments like reflections and open questioning. The use of a wide variety of assessments is aligned with the learning and assessment goals of the *more aligned* teachers, as well as their views of their students. Because these teachers wanted to prepare their students for life outside of school, their assessment and learning goals went beyond content and knowledge to include geographical attitudes and dispositions. To this end, they used a range of assessment tasks including reflections to assess skills, values and dispositions. Their employ of a wide range of assessment types resonates with the syllabus, which calls for teachers to assess student learning using a variety of assessment types, including portfolio, fieldwork, and oral presentation (Curriculum Planning and Development Division, 2005).

The *moderately aligned* teachers relied on both formal (e.g., the tests that Amanda and Miki submitted) and informal (e.g., the in-class worksheets from Jiajia and Margaret) assessments. The *less aligned* teachers used the smallest range of assessments, despite the syllabus urging teachers to use a wide range of assessment types and formats. They spoke most frequently of formal assessments, and these assessments adhered closely to the test blueprint provided in the syllabus. The use of formal assessments and worksheets that were based on the test blueprint are aligned with the assessment and learning goals of these two groups of teachers. The *moderately* and *less aligned* teachers focused on assessing facts and content because they perceived their lower secondary students to be lacking in basic geographical knowledge. On the whole, there was most variety in assessment practices among the *more aligned* teachers.

Persistence. The second pattern of persistence refers to the continued attention paid to assessing discrete and unrelated content knowledge in spite of the policy's focus on understanding, application and other higher-order skills. The data presented in Chapters 4 and 5 indicated that Singapore teachers have continued to pay attention to assessing content knowledge. The *moderately* and *less aligned* teachers' comments about the assessments they submitted pointed to the unrelenting emphasis placed on the recall of facts and knowledge without requiring students to make any real world application or demonstrate understanding of such knowledge. This focus was most evidently seen in the *moderately* and *less aligned* teachers' comments about their learning goals and objectives, which indicated that facts and knowledge from the textbook were necessary for performing well in the midyear and the end-of-the-year examinations. They also saw the preparation of lower secondary students for the upper secondary examinations as part of their learning goals and objectives. The *moderately* and *less aligned* teachers focused on assessing facts and content because they viewed their lower

secondary students as lacking knowledge. While they did not dispute the need for developing and assessing higher-order skills, they preferred to evaluate these skills at the upper secondary level when, in their view, students had developed sufficient foundational knowledge and skills. It was not that the *more aligned* teachers did not value facts and content. Rather, their assessment practices did not require their students to repeat learned facts because they wanted their students to be able to apply and understand what was taught. As Totoro said, given the easy access to and availability of information in a globalized age, the goal was to assess students' ability to gather, analyze and synthesize information and apply knowledge to real world contexts.

The pattern of persistence was also exemplified in the way the teachers approached formative assessment. Instead of doing more guiding, facilitating, and modelling as envisioned in TSLN, the *moderately* and *less aligned* teachers continued to "tell" their students the right answers or require students to diligently copy model answers. Even among the *more aligned* teachers, the practice of inviting students to make sense of their errors and misconceptions was not frequently carried out, nor offered to all students. In sum, the continued emphasis on assessing factual content and the lack of opportunities for students to play more active roles in the classroom suggest that despite TSLN's call for teachers to nurture thinking students, the majority of teachers did not enact classroom assessments that resonated with the policy.

Change. The third pattern of change indicated that some teachers were paying more attention to assessing higher-order skills as envisioned in the policy and to using a range of assessment types as recommended in the revised syllabus. This is to say, their assessment practices reflected the TLLM tenets to use more qualitative and formative assessing and less quantitative and summative testing. They were also addressing the area of understanding rather than the sole emphasis on knowledge. This was evident from the way the *more aligned* teachers

spoke about their assessment and learning goals and their views of their students. These teachers focused on assessing not just geographical content, but also the skills and values encapsulated in this discipline. They also used prompts that required students to discuss their responses from multiple perspectives, and consequently, had mark schemes that allowed for a range of plausible responses. Furthermore, they used a range of assessment types; in fact, only the two *more aligned* teachers submitted assessments that took the form of independent projects and research. The characteristics of their assessments resonated with Newmann and Associates' (1996) AIW criteria. Although the *more aligned* teachers also taught lower secondary students, unlike their *moderately* and *less aligned* colleagues, they did not perceive these students as lacking in geographical knowledge and consequently, as a bottleneck to their learning and assessment goals. As Harry pointed out, it was possible to provide students with challenging tasks. In his view, teachers just had to pose the appropriate issues and students would be able to draw from their experiences when completing the task.

While the *moderately aligned* teachers spoke about the need to prepare students for the future, their assessments did not reflect these goals. In fact, apart from an occasional prompt here and there, the *moderately aligned* teachers mostly assessed discrete pieces of factual content knowledge. Some *moderately aligned* teachers emphasized the teaching and assessing of some higher-order skills like extended writing. However, this emphasis was because the assessments for the upper secondary examinations had changed and thus, they valued using open-ended prompts as a form of examination preparation rather than because of the importance of teaching students to communicate their learning through extended writing. To this end, while there was some alignment towards the emphasis on higher-order skills, only the goals of the *more aligned* teachers reflected the policy intent. The goals of the *moderately aligned* teachers detracted

somewhat from the policy goal. The pattern of change towards emphasizing higher-order skills and using a variety of assessment formats and types was not evident in the *less aligned* teachers' practices. The two teachers in this group used assessments that did not address higher-order skills in the proportion stipulated by the syllabus. In fact, the proportion of higher-order skills to content and facts in Maryanne's assessments was less than that stipulated in the assessment objectives.

Overall, the *more aligned* teachers enacted assessment practices that reflected variety in terms of assessment types and forms; some change towards paying attention to some higher-order skills, but also some persistence in assessing discrete and unrelated items of content knowledge. The variety was evident in their use of a larger range of assessment types than their colleagues. The pattern of change towards addressing 21st century skills was evident in the two teachers' comments about wanting to equip their students with life skills and in the attention they paid to assessing application and understanding of concepts and content. Overall, the comments and assessments these teachers submitted showed little preoccupation with the recall of factual knowledge and content. The pattern of persistence in adopting practices TLLM was encouraging teachers to do less of was manifested in the way these two teachers continued with teacher-centered classroom assessment, providing minimal opportunities for students to be actively involved in assessment. However, the extent of the pattern of persistence was lowest for the *more aligned* teachers.

Among the *moderately aligned* teachers, there was evidence of some variety in assessment practices in terms of how some teachers spoke about marking students' tasks without assigning scores and about asking students to express their responses through drawings. There was modest change in practices towards assessing the outcomes envisioned in TSLN. This was

reflected in the assessments teachers like Jiajia and Margaret submitted which incorporated occasional prompts to assess students' dispositions and attitudes. The pattern of persistence in preparing students for a life of tests was evident in the teachers' focus on passing examinations, and in their comments about aligning their curriculum and assessment to dovetail with the requirements for the high-stakes upper secondary examinations.

Finally, the *less aligned* teachers' practices showed no evidence of variety in terms of assessment formats or types, or of change towards assessing the outcomes envisaged in the policy. Their practices did, however, reveal patterns of persistence in emphasizing examination and test preparation. This was because they frequently spoke about testing and retesting students to ensure that they were sufficiently familiar with various permutations of assessment prompts. These teachers assessed only facts and content in ways that the policy is asking teachers to enact less of. For formative assessment, the *less aligned* teachers emphasized test-taking strategies and provided students with model answers rather than encouraged students to reflect on their errors.

Together, the combination of the three patterns of change, variety, and persistence suggests that, overall, teachers' assessment practices largely diverged from the policy intent. The pattern of persistence in emphasizing quantitative and summative testing which TLLM was encouraging teachers to use *less* of was strongly manifested in the practices in all three groups of teachers. Furthermore, six of the eight teachers continued to lead classroom discourse during formative assessment resulting in students waiting passively for the right answers and solutions. This departs from the policy vision which is calling for teachers to do less "telling" and to do more facilitating and guiding. While the *more aligned* teachers would, from time to time, invite students to work out their misconceptions, these occasions were not frequent, and sometimes only accorded to higher-ability classes. As a result, teachers' views of student ability, as well as

their perceptions of teaching, learning, and assessment continued to resonate with practices that TLLM was calling for teachers to use *less* of.

In comparison, the patterns of change towards assessing skills that prepare students for the 21st century and also using a variety assessment practices were not strongly reflected in the practices of the majority of teachers. Despite the fact that the revised geography syllabus had called for teachers to use a variety of assessments to assess students' acquisition of geographical skills, knowledge, and values, the pattern of variety in using different assessment types and formats was not evident in the *moderately* and *less aligned* teachers' practices. Only the two *more aligned* teachers submitted assessments that had a variety of formats.

Furthermore, the pattern of change towards assessing cognitive skills other than discrete facts and content was not widely practiced by the *moderately* and *less aligned* teachers. Although the policy and syllabus emphasized the teaching and assessing of higher-order skills, only the two *more aligned* teachers spoke about assessing higher-order skills in a way that reflected the policy intent. Among the *moderately aligned* teachers, the change towards the assessment of higher-order skills as envisioned in the policy was largely to prepare students for upper secondary examinations, rather than to equip students with 21st century skills. Thus, the *moderately aligned* teachers' assessment of skills such as elaborated written communication at the lower secondary level was not congruent with the policy intent, but was enacted to prepare students for high-stakes examinations which they would undertake in two to three years' time.

As a result, the combined patterns of variety, change and persistence indicate that a decade and a half after the launch of TSLN, there was just "incremental change" in assessment practices (Cuban, 1993) in the sense that teachers merely incorporated small changes, such as including a higher-order prompt here and there, within their existing assessment practices. There

were no changes in the learning culture in the classroom nor were students encouraged to play more active roles in the learning process.

Learning versus achievement

The TSLN and TLLM slogans contain 'learning' and 'learn' respectively. A "learning nation" in TSLN is about making learning the national culture and about recognizing that education lies on a continuum beginning with pre-school and continuing throughout life (C. T. Goh, 1997). Education, therefore, continues beyond schools and educational institutions (C. T. Goh, 1997). This view of continuous learning resonates with the idea of a "learning society" in which learning is part of daily life rather than being merely confined to schools and colleges (C. Watkins, 2011, p. 3).

Views about learning can be grouped into three broad categories, namely *Learning is being taught*, *Learning is individual sense-making*, and *Learning is building knowledge as part of doing things with others* (C. Watkins, 2011). In *Learning is being taught*, learning is about the mind being filled, like a container, and assessment involves checking if the learning is there (C. Watkins, 2011). This view of learning resonates with the behaviorist (James, 2008) or the quantitative tradition (Biggs, 1995, 1996a) as teaching and assessment are based on transmission from teacher to learner and the aim is the acquisition of facts, skills and behavioral objectives (Biggs, 1996a). In the quantitative tradition, the aggregate score students acquire or achieve indicates competence in what is learned (Biggs, 1995). Such assessment is "first generation assessment practice" in which assessment focuses on "what is taught," and learning is assumed to have taken place when the knowledge is "retained" (James, 2008, p. 21).

The second conception of learning, *Learning is individual sense-making*, emphasizes the importance of learners taking charge of their learning. In other words this view of learning values

self-directed learners (C. Watkins, 2011). In this conception, learning is an active process of sense-making and is affected by the use to which knowledge is put. Thus, learning occurs when learners are be able to explain ideas and concepts to themselves or to others (C. Watkins, 2011). This conception of learning draws on cognitive constructivist views of learning and conceives of learning as being determined by what happens in the learners' heads (James, 2008).

The third category, *Learning as building knowledge as part doing things with others*, views learning as being constructed as part of a social activity, especially through dialogue (C. Watkins, 2011). This view draws on socio-cultural theory which conceives of learning as involving "thought and action in the context" of learning. In addition, learning is the "interactions among these phenomena" (James, 2008, p. 29).

Together, these latter two views of learning reflect the characteristics of learning within the qualitative tradition which conceive of students actively interpreting and incorporating new material with their prior knowledge (Biggs, 1995, 1996a). The assessment of *Learning as individual sense-making* involves assessing the individual (James, 2008) but the focus is on students' ability to understand and solve problems (Biggs, 1995; James, 2008), rather than to recall and reproduce knowledge. The assessment of *Learning as building knowledge as part of doing things with others* requires assessment to be carried out as learning takes place and not after learning has been completed (James, 2008). In addition, assessment is conceived as being carried out by the community of learners rather than by external assessors (James, 2008; Shepard, 2000). There is also a role for peer and self-assessment (James, 2008; Shepard, 2000). Mary James (2008) terms these types of assessments "second generation" (p. 25) and "third generation" (p. 29) respectively.

In calling for teachers to do more *qualitative* assessing, TLLM is encouraging the second and third conceptions of learning. However, based on the assessments submitted and the teachers' comments about their formative assessment practices, I suggest that for the majority of the teachers, the emphasis appears to be achievement rather than learning. For the teachers who focused on achievement, learning simply meant 'being taught,' as they wanted students to reproduce what they had been told in class. Comparatively, it was evident that the *more aligned* teachers, as well as Margaret, focused on learning as their discussion of student work indicated that they looked at what students were able to do rather than on simply emphasizing the marks that students obtained. These teachers' attention to learning was also seen in the way they discussed the quality of students' responses such as the strengths of the response, the change in the quality of work over the school year, and the quality of the thinking as shown in the response.

Focus on achievement. The *moderately* and *less aligned* teachers placed an overwhelming emphasis on achievement, or the accumulation of marks. This was seen in the way these teachers exposed their students to permutations of how a concept or topic could be assessed, zoomed in to focus on grades and marks when discussing their analyses of student work after marking, subjected their students to numerous tests and retests, and focused on examination and test strategies. This pattern of practices suggests that these two groups of teachers continue focusing on a *life of tests*. In comparison, the *more aligned* teachers were neither preoccupied with achievement nor with the accumulation of marks. Rather, as Totoro pointed out, the emphasis was on what students showed they were capable of doing, and subsequently, to design tasks that build on these skills.

The teachers' focus on achievement was also evident in the manner in which they enacted formative assessment. For instance, they emphasized the managerial aspects of feedback, such

as telling students to be mindful of the time (e.g., Amanda), to use and apply success strategies to respond to questions (e.g., Jiajia), and to take down model responses (e.g., James and Maryanne). By adopting such approaches, the teachers "reduce thinking" to "techniques, strategies, mental processes, procedures, or 'correct' examination answers (Koh, et al., 2012, p. 141) which run counter to the TSLN intent. This is because higher-order thinking cannot simply be reduced to a set of procedures and strategies like Jiajia's three-step approach to interpreting geographical data. Rather, if teaching and assessing are to be aligned to the TSLN vision, then beyond these techniques is the need to develop the "dispositions or habits of mind of a higher-order thinker"(Koh, et al., 2012, p. 141), as shown in the practices of Harry and Totoro. Therefore, while the *moderately* and *less aligned* teachers said that they taught and assessed higher-order thinking, their practices indicated that they merely provided their students with numerous worksheets, exercises, and tests to practice thinking skills (Koh, et al., 2012). While such approaches may enable students to succeed in tests and examinations, Koh et al. (2012, p. 141) doubt if *thinking schools*, which require a "culture of critical reflection … and real-life problem solving within and outside the classroom" can be achieved.

Teachers who overemphasize correct responses, unintentionally bring about less spontaneity and lower levels of engagement from their students (Koh, et al., 2012). As a result, students merely sit and absorb facts and knowledge through passive listening (Koh, et al., 2012). This is perhaps why the *moderately* and *less aligned* teachers often lamented that despite re-teaching and numerous practice exercises, their students could not remember facts or strategies, and continued to make the same mistakes. In Miki's experience, sometimes students did not even bother to study for the test, and she also wondered why her students were reluctant to raise their hands to answer her questions.

The research on formative assessment (e.g., Riggan & Oláh, 2011) has reported that teachers tend to focus on errors rather than address misconceptions in student learning. In Singapore, the relentless drive to ensure that students know the right answers exists because of the examination-oriented culture (Koh, et al., 2012). As evident from the teachers' comments, there is a responsibility to prepare students to perform well in formal assessments and to maintain the percentage of passes and grades (Koh & Luke, 2009). To this end, the *moderately* and *less aligned* teachers spent a significant amount of time and energy on drill and practice in order to ensure that their students could perform well in the examinations.

Focus on learning. Some teachers' practices did focus on learning in the way espoused by TSLN. Harry and Totoro's assessments required students to apply what they learned as well as to demonstrate understanding. Their approaches to learning resonate with Watkins' (2011) views of learning, namely *Learning as being taught* and *Learning as individual senses-making.* These strategies are manifested in the way Harry tasked his students to write reflections after each assessment. This is also the reason why Harry did not want his students to merely copy model answers as is the practice of the *moderately* and *less aligned* teachers. He preferred that his students work through the responses themselves. As Totoro pointed out, when students work out the misconceptions themselves, it is "more powerful" than when the teacher supplies the response. Thus, Totoro capitalized on the currency of the farm experience to ask her students to discuss and identify good interview prompts and skills. In so doing, the *more aligned* teachers helped their students to understand the purpose of the activities and tasks, and thus enabled students to remember what they learned.

The formative assessment strategies Harry and Totoro employed were more successful than those of their *moderately* and *less aligned* colleagues in helping students learn. Compared

to those two groups of teachers who bemoaned that their students did not remember all the nagging and advice, the *more aligned* teachers were heartened that their students were engaged in the subject and were able to demonstrate application and understanding of what had been taught. Harry's students would point out the relevance between concepts they were taught in class with what they read in the newspapers, and Tototo and her colleagues were surprised that her students could "remember" the experiences on the farm. The spontaneously penned reflection by Totoro's Secondary 1 student asking for more assignments like the fieldwork task (see Chapter 5) illustrates how students can be engaged in learning when they are assigned meaningful and challenging tasks. The strategies Harry and Totoro used involved a "deep approach" to learning because there is a focus on the meaning underlying what is to be learned (Marton, et al., 1996, p. 69). For this reason, in terms of their learning goals and assessment objectives, Harry and Totoro did not just focus on geographical content, but also helped their students to develop geographical skills, attitudes and values. In comparison, the *moderately* and *less aligned* teachers' approaches merely led to "surface learning" as they only emphasized the tasks or the content to be learned and not the underlying purpose (Marton, et al., 1996), as manifested in the regurgitation of content that teachers like James expected of their students.

On the whole, while the policy envisions teachers enhancing student learning by teaching less, the assessment practices suggest strongly that the majority of the teachers privilege achievement over learning. Apart from the two *more aligned* teachers, the other teachers conducted formative assessment within the quantitative tradition as their intent was to help students accumulate more marks. It is understandable that the teachers are committed to help their students do well, as Harry said that teachers had to "help those who don't score the perfect score." However, the *moderately* and *less aligned* teachers' approaches to enhancing learning

did not mirror the policy intent. Rather than help students understand errors and misconceptions, the *moderately* and *less aligned* teachers merely dispensed correct and desired answers, and repeated and recapitulated strategies that students need to use in order to do well. Only Harry provided opportunities for his students to revise their work after giving them feedback. It was through the revised drafts that Harry and his students were able to determine if the gaps in learning had been closed. And only Totoro tasked her students to compare their own responses with exemplars she provided. The aim of this formative assessment approach was for students themselves to arrive at an understanding of where they went wrong. The other teachers merely treated students as empty "vessels" to be filled by making them take down model answers and imbibe test-taking strategies.

Overall, the theme of *learning versus achievement* suggests that for classroom assessment practices to be aligned to TSLN, teachers need to understand the meaning of 'learning' as espoused in the policy, and to have the ability to bring about this learning. The current view held by the *moderately* and *less aligned* teachers is that learning is the acquisition of knowledge. This mindset may need to shift towards learning to acquire knowledge. This practice was adopted by the *more aligned* teachers, and manifested by the research tasks Harry and Totoro used which required students to look up multiple data sources, analyze them, and to construct a narrative based on a given geographical issue.

To this end, the teachers' focus on achievement over learning through the use of teacher-centered teaching strategies points to a policy-practice disconnect in which the teachers are approaching assessment from a perspective that is different from that undergirding the policy. While the TLLM tenet urged teachers to do *less* "telling" and *more* facilitating, it envisioned that students would participate more actively in the learning process. It is clear also that the policy

envisages that students engage in more active and independent learning. This suggests that the policy is aligned to socio-cultural theory, or the conception that *Learning is building knowledge by doing things with others*. However, the analyses indicated that the teachers did not create many opportunities for student participation. In fact, the majority of the teachers continued to privilege the use of teacher-centered practices aligned with the quantitative tradition over the qualitative tradition. Even though the *more aligned* teachers would, when possible, ask students to make sense of their mistakes and misconceptions, they too did not engage their students in peer or self-assessment in their classrooms. The implications of this policy-practice gap for practice in school and for policy will be discussed in Chapter 7.

Explaining teachers' assessment practices

This chapter used inductive and deductive coding to analyze the teachers' interview comments to understand the classroom assessment practices they used. Deductive coding enabled the description and analyses of the extent to which the teachers' classroom assessment practices were aligned to constructivist assessment, and by extension, to the TSLN and TLLM policy intents. However, as constructivism is a descriptive theory of learning (Richardson, 1997), inductive coding was used to elicit emergent themes to explain the teachers' assessment practices. Three factors emerged from the inductive coding, namely the degree of alignment of professional perspectives to policy, the supportive or less supportive nature of school policy, and opportunities for engaging in professional learning and collaboration.

The teachers' professional perspectives about their students' abilities and about their roles in teaching and learning influenced the nature of their assessments, and in particular, the way they assessed higher-order skills. The teachers who viewed their students as having the ability to develop, and as being able to work on challenging tasks that were scaffolded had

assessment practices that were aligned to the policy. In addition, these teachers saw themselves as facilitators of learning, and hence, their formative assessments practices guided students to close their learning gaps. In comparison, teachers who viewed their students from a deficit perspective, and who saw themselves as examiners tended to use assessments that focused on eliciting facts, concepts, and definitions instead of assessing understanding and application. Their formative assessment practices served to dispense correct answers and ensure students knew of different success strategies to achieve more marks.

The nature and extent of alignment to policy or innovation is contingent on the attitudes and values of teachers (Priestley, 2005; Priestley & Sime, 2005). Since the decisions teachers make are driven and guided by their views of teaching, the way they view themselves as professionals and their perceptions of their students, classroom change is therefore a highly personal experience for teachers (Coffey, et al., 2005). Of the three factors that provide plausible explanations for teachers' assessment practices, there was the most divergence in the professional perspectives teachers held, with five teachers having views of teaching and assessment that resonate with the "20th century dominant paradigm" of behaviorism and efficiency (Shepard, 2000) and the other three teachers speaking of perspectives that echo the "emergent paradigm" of constructivism (Shepard, 2000). This suggests that more teachers held views of teaching and assessment that diverged from the TSLN and TLLM intent. Broadly, teachers whose assessment practices reflected the TSLN intent articulated professional perspectives of students and of their roles in teaching and learning that echoed constructivist theory. In particular, the *more aligned* teachers as well as Margaret saw their roles as facilitators of learning. In guiding students to make sense of misconceptions rather than dispensing model answers, these teachers engaged their students in meta-cognition, a feature of cognitive and

constructivist learning theory (Shepard, 2000; Torrance & Pryor, 2001). Additionally, these teachers saw their students as having the ability to learn, and thus provided tasks that were challenging, yet scaffolded to ensure that students could complete the assignment. In short, these teachers' approaches to teaching and assessment were consistent with the theory and principles underpinning the TLLM tenets. Guided by these perspectives, the teachers wanted their assessments to address a range of learning, including reflection and research, as well as students being able to capture their understanding of concepts through drawings.

In comparison, teachers whose assessment practices diverged from the TSLN intent saw themselves as supervisors of learning as well as examiners, practices that echo behaviorist theory. They also saw their students as being deficient in basic content. Because these teachers held such perspectives, they assessed their students' ability to recall and reproduce discrete and unrelated content, providing more challenging tasks only after students demonstrated their mastery in basic facts. In addition, as the *less aligned* teachers and some *moderately aligned* teachers (e.g., Miki and Amanda) did not see it necessary to provide challenging tasks to lower secondary students, their assessment tasks did not require students to engage in higher-order analyses, evaluation, or application of content to a real world context. Instead, many of the tasks they presented their students merely had "closed" questions with binary right-wrong responses (Torrance & Pryor, 2001, p. 617). Teachers like James even demanded formulaic responses from their students for opened-ended questions. When enacting formative assessment, these teachers failed to recognize that learning involves a process of active sense-making rather than the passive absorption of knowledge (James & Lewis, 2012). Consequently, their formative assessment practices focused on "contrasting errors with correct responses" (Torrance & Pryor, 2001, p. 617). These practices, rather than reflect TLLM's underlying philosophy, resonate with

characteristics of behaviorist and hereditrian theories, in which learning is presented in atomized bits that are sequenced and arranged in a hierarchy (Shepard, 2000). Therefore, the tests and retests are merely the "hurdle" that students need to cross to indicate that they are competent and ready for the next topic or concept (A. Hargreaves, et al., 2002, p. 76).

Second, policies at the school or meso level affect the nature of the teachers' assessment practices. Teachers who had a wide repertoire of assessment approaches, especially those who used alternative assessments like independent research, worked in supportive schools. Such schools provided administrative, financial, resourcing and personnel support that enabled the teachers to assess their students differently. In particular, for teachers like Totoro to be able to conduct an assessment based on a fieldwork experience required the school to free up colleagues' schedules so that they could also accompany students to the farm. Therefore, the presence of a supportive and encouraging school environment is necessary for bringing about assessment change. In particular, school culture aids in teachers' willingness to take risks, and to question and reflect on practices (Jones & Moreland, 2005). Such school support was evident in Totoro's school, which led her to speak about the need for teachers to be creative and to learn to deal with ambiguity by themselves when they tried out new types assessment tasks.

Teachers' assessment practices were affected by their schools' administrative, structural, and resourcing policies. Although some teachers may wish to be creative and innovative, if their administrators do not lend support by providing time and resources, these teachers' assessment practices will be less reflective of the macro policy intent. Teachers' reluctance or inability to adopt TLLM-envisioned assessment practices is exacerbated when schools adopt harsh policies on teacher appraisal and ranking, which can ultimately reduce creativity and risk taking (Liew, 2008). This resulted in teachers like Maryanne continuing to adopt assessment practices that

merely focused on examination preparation. The consequence of such appraisal policies was that teachers continued using strategies that had worked for them in the past. To this end, to show their supervisors that they were working hard to prepare their students, the teachers' assessments mirrored the test blue print closely. They also focused on drill and practice to ensure students were nonplussed by prompts used, and hence, cruise through the examinations. In the case of the two *less* aligned teachers, the consequences of their schools' use of the MSG value to appraise their effectiveness were undoubtedly high-stakes for the following reasons. First, their schools evaluated them solely on student results instead of assessing their contribution to students' overall academic and character development as manifested under the other 16 competencies in the teacher evaluation document (OECD, 2011). Second, these teachers were teaching lower secondary students; a non-key stage level. Students' grades at this level were not being used for high-stakes purposes. Yet the schools Maryanne and James worked in were ranking and appraising teachers teaching a non-key stage level intended to initiate students into secondary education. Finally, school support needs to be provided and sustained in order for the changes in practices to commence and deepen (Jones & Moreland, 2005). However, there was little support for the two *less aligned* teachers. Comparatively, Totoro's and Harry's experiences show the value of sustained administrative support in timetabling and in promoting a culture of inter- and intra-departmental collaboration which enabled them to construct research tasks and fieldwork. In the example of Harry, there was sustained support from the school over a three-year period for the inter-disciplinary task he and his colleague designed.

Third, the majority of the teachers—with the exception of the two *less aligned* teachers—found professional learning an important way to improve their classroom assessment practices. However, more than just attending workshops, conferences and sharing sessions, teachers had to

experiment with the ideas they gleaned at these sessions and apply them in their assessments. In addition, opportunities to engage in cross-departmental collaborations with colleagues also influenced the nature of teachers' assessments. An important outcome of professional learning and collaboration is that teachers discuss and explore assessment issues rather than focus solely on administrative procedures related to testing. Another significant consequence was the establishment of learning communities in the school, which according to Jiajia was beneficial because when colleagues learned something at a professional development platform, they returned and shared with their peers. One aspect related to professional learning is collaboration with colleagues, both formal and informal, which provided opportunities for teachers to construct and use innovative assessments. Through close collaboration with colleagues, Harry and Totoro were able to try out interdisciplinary assessments and fieldwork assignments.

Participating in professional learning and collaboration are ways to develop teachers' assessment practices and guide them to align their approaches to the policy intent. However, professional learning goes beyond attending centrally organized workshops and seminars. More importantly, in terms of TSLN, the nature of these professional learning sessions needs to reflect the policy intent.

In this study, the school-based learning platforms that were initiated and organized by the *more and moderately aligned* teachers resonated with TSLN's intent of driving change and reform from the school or grassroots level rather than from the top (Shanmugaratnam, 2005a). The move towards having teachers own and drive professional learning began in the early TSLN years, when the Teachers Network—a department within the MOE—supported teacher-initiated activities such as reflections and peer collaboration in order to improve teaching and learning practices (Hairon, 2008). At the time of this study some ten years later, teachers themselves

were initiating and participating in learning, reflection and collaboration in professional learning communities (PLCs) within their schools. These teacher-initiated professional learning sessions have been a mainstay in efforts to build teachers' capacity in Singapore (Hairon, 2008) and serve as platforms for teachers to learn from one another with the aim of reviewing and improving teaching and learning (Mourshed, et al., 2010).

School-based professional learning echoes the socio-cultural theory of learning in which *Learning is building knowledge as part of doing things with others*. The teachers' enthusiastic comments about learning from and with colleagues in their schools' professional learning teams indicated that at the school-level, learning within the school-based community was a powerful way to generate conversations among colleagues about issues and practices related to assessment, teaching and learning. This form of professional learning is closely aligned with socio-constructivist theories where learning takes place in a communal setting and is distributed across the community (James, 2008). School-based professional learning manifests the notion of *Learning is building knowledge as part of doing things with others*, and departs from the *Learning is being taught* perspective of centrally-organized workshops in which, for example, James listened to examiners who talked about the examination and marking process.

Many lenses may be used to study classroom assessment (Black & Wiliam, 2012b). This dissertation has drawn on constructivist theory to analyze and interpret the teachers' assessment practices. Based on its analyses, the teachers whose practices reflected the principles and features of constructivist theories were guided by perspectives of teaching, learning, and of their students that also reflected the policy. Conversely, teachers whose practices were in dissonance with the policy were guided by perspectives of teaching and learning that resonated closely with behaviorist theories, an older tradition in education (Biggs, 1996a), and one whose practices

TLLM is urging teachers to use less of. To this end, one implication from the findings that will be discussed in Chapter 7 is the need to articulate more explicitly the philosophies and principles underlying the reform so that teachers might become more aware of the spirit and intent of the policy. More specifically, one necessary step for ensuring greater congruence with the policy is to engage teachers, and to help them understand the policy in relation to their values, beliefs and perspectives.

Together, the three themes of professional perspectives, school policy, and professional learning provide plausible explanations for the variations in classroom practices by the teachers in this study. Based on the teachers' experiences, the most significant factor affecting teachers' assessment practices appears to be *policies in school*. This is because the degree and extent of support that schools give to teachers' curricular plans affects the nature of assessments that they are able to carry out. The eight teachers in the study experienced different degrees of school support. The *more aligned* teachers had the most school support while the *less aligned* teachers had the least. The support received by the *more aligned* teachers enabled them to conduct inter-disciplinary assessments and fieldwork tasks. The extent to which school policies influenced the *moderately aligned* teachers' assessment practices was mixed: some teachers could improve their knowledge of classroom assessment by participating in professional learning communities while others had to adhere to rigid departmental plans which made it difficult for them to better customize their assessment practices to students' needs. Furthermore, when schools continued to evaluate teachers using only students' results, teachers responded by adopting practices that had worked well for them in the past and were reluctant to try or adopt new assessment practices. Thus, administrative processes, structures, environment, and culture in schools can either facilitate or hinder change in teachers' assessment practices.

Teachers can and do embrace different views and perspectives of teaching, learning, and assessment, of students, and of their professional roles. But how they enact these in the classroom is strongly dependent on the schools they work in, as well as on those schools' policies. Thus, even if teachers hold views that are in concordance with the TSLN intent, the way the school is run and the nature of the culture in the school can directly and indirectly influence the form that their classroom assessment practices take. Similarly, while the majority of teachers in the study found professional learning to be beneficial, the way they applied and translated what they gleaned from workshops, conferences, and sharing sessions depended largely on the nature and degree of support that was provided by their schools.

How can teachers better understand the policy? How can schools support teachers to enact assessment practices that are aligned to the policy? How can school policies be aligned to the intent of the policy? These questions and their associated implications will be discussed in Chapter 7.

CHAPTER 7: IMPLICATIONS AND CONCLUSION

Introduction

Preparing students to be ready for the 21st century involves developing their ability to work independently, to devise strategies in order to solve problems, and to be able to apply higher-order skills such as critiquing, synthesizing and interpreting a variety of data, communicating effectively, and solving problems (Darling-Hammond, 2010; C. T. Goh, 1997; Kay, 2010; Luke & Hogan, 2006). Since 1997, Singapore's education system has focused on equipping students with these skills in order that they are able contribute to the economy and society. This process was initiated by articulating a clear vision of education, and by reviewing and revising syllabuses and assessments to focus on the intended 21st century skills.

My personal interest in this topic arises from my experience as a student and as an educator in Singapore. I was educated, and started teaching during the efficiency-driven phase (1979-1996). Shortly after I commenced teaching, TSLN was introduced, and I have been intrigued by its focus on higher-order thinking skills, and its compelling vision to prepare students to live in and contribute to the 21st century society. I am particularly interested in teacher assessments having been a classroom teacher myself, and having experienced the phenomenal pressure of preparing cohort after cohort of students for the high-stakes pre-university examinations. In light of this, I was interested in examining how teachers enacted classroom assessments that reflected the TLLM tenets, since the research reports that the way teachers assess student learning appears to be resistant to change (Tierney, 2006).

This chapter provides a meta-inference of the findings reported in Chapters 4, 5 and 6. The mixed methods discourse calls for an overarching or meta-inference that integrates interpretations from the two aspects of the study to address the entire study (Onewuegbuzie &

Teddlie, 2003; Tashakkori & Teddlie, 2008). To this end, this chapter applies a cross-over strategy of "warranted assertion analysis" (Onwuegbuzie & Combs, 2010), which involves the use of all data sources to arrive at meta-inferences to answer the overarching research question. Specifically, the meta-inference addresses the overarching research question:

> **Under an educational policy that emphasizes the preparation of students for "the test of life" instead of a "life of tests," (MOE [Bluesky], 2005), how do Singapore geography teachers elicit and enhance student learning through the ways they use classroom assessment?**

This chapter first summarizes and synthesizes the findings from Chapters 4, 5 and 6, and then proceeds to present the meta-inference that integrates the inferences from the three chapters. It concludes by examining the limitations of the study and discusses implications for policy, research and teacher educators, as well as for practice in schools.

Teachers' Classroom Assessment practices

Making changes to teachers' classroom assessments represents major paradigm shifts in thinking about learning, schools, and teaching (A. Hargreaves, et al., 2002). To realize the TSLN vision, Singapore's Ministry of Education, through the TLLM tenets, urged teachers to do *more* "qualitative and formative assessing" and *less* "quantitative and summative assessing." The TLLM tenets also remind educators that education's purpose is to prepare students for *the test of life*, and not to subject students to a *life of tests*. These tenets signal a shift towards constructivist learning theories, or what Shepard (2000, p. 5) refers to as the "emergent paradigm," and call for teachers to reduce the emphasis on behaviorist learning approaches, or the "20th century dominant paradigm" (Shepard, 2000, p. 5). While the TLLM tenets do not make explicit reference to terminology from the different learning theories, there are clear indications of the policy's emphasis on constructivist theories. Finally, in advocating a more of/less of approach, TLLM does not call on teachers to adopt pendulum shifts in their teaching

approaches, but rather to adopt and use a wider repertoire of pedagogies and assessment practices so that they can equip students with the skills needed for the 21st century.

To examine the way teachers elicited and enhanced student learning through their classroom assessments, this dissertation used a mixed methods approach to analyze assessment practices over time, both at the macro (national) and micro (classroom) levels. The macro (national) data was analyzed using the *z*-score and documentary analyses. The micro (classroom) level data was analyzed quantitatively using descriptive statistics and paired *t* tests, and qualitatively using theoretical and open coding (Maxwell, 2013).

The responses of teachers teaching a representative sample of students drawn from five TIMSS cycles provided a macro (national) picture of practices over time (from 1995 to 2011), while eight teachers' classroom assessments administered at the time this study was conducted presented a micro (classroom) pattern of current practices. The quality of the assessments teachers presented to their students was analyzed using Newmann and Associates' (1996) AIW criteria which served as indicators of the higher-order skills that TSLN wants students to acquire and demonstrate. Drawing on the assessments the teachers submitted, and based on the interview comments, this dissertation analyzed the assessments teachers used to elicit and enhance student learning, and sought to understand why they enacted these practices. While each TIMSS cycle surveyed teacher assessment and student learning in four science components—physics, chemistry, biology, and earth science—the classroom level data focused solely on the earth science (or geography) component in order to analyze in greater depth assessment practices in this content domain.

Nature of classroom assessment

Together, the macro and micro findings suggest that there are three patterns of assessment practices, namely, *variety*, *change and persistence*. The pattern of *variety* refers to the use of a range of assessment types, including formal and informal assessments, and objective and constructed-response questions. The pattern of *change* points to the use of assessment practices that are aligned to the TLLM tenets, such as more qualitative and formative assessing and less quantitative and summative testing. Third, the pattern of *persistence* reflects the use of assessment practices that the policy is encouraging teachers to use less of. The patterns of *change* and *variety* indicate alignment with the policy while the pattern of *persistence* suggests continuity and consistency in the teachers' practices.

Variety. For the pattern of *variety*, based on the analyses, there was convergence between the micro and macro data. Specifically, at the macro level—as presented in Chapter 4—Singapore students had teachers who used a range of assessment types. This indicates that at a national level, Singapore students had teachers who did not exclusively rely on one or two types of assessment. Rather, they employed a combination of practices. This mixture of assessment types indicates that teachers used a combination of formal and informal assessments, and traditional (i.e., examinations) and performance assessments (e.g., research projects) to obtain an idea of student learning (Ohlsen, 2007).

The analyses of the micro data showed alignment with the macro data. From the assessments which the eight teachers submitted to this study, there was a *variety* of formal and informal assessment types. The former encompassed projects, common tests, class tests, and school examinations while the latter included in-class tasks and practices embedded within pedagogy and instruction, such as observing and questioning students. Informal assessment

practices are those teachers used to gauge the extent of student learning and understanding and are associated with good instruction (Nitko & Brookhart, 2011). The micro data also indicated that there were variations among classrooms as some teachers were using a larger variety of assessments than their colleagues.

The similar macro and micro level findings for the pattern of *variety* suggest that there was some conformity to the policy intent in that, despite the need to prepare students for school and national examinations, teachers did not exclude the use of other assessment types like performance assessments to obtain a picture of student learning.

Change. The second macro pattern is *change* which was represented by assessment practices that reflected the policy intent. At the macro level, this pattern emerged from the reports of students' teachers to the TIMSS questionnaire items on the frequency of testing, and the emphases on assessing the cognitive domains of *knowing*, *applying*, and *reasoning*. The macro data indicated a reduction in the frequency of testing over the period from 2003 and 2007—the later TSLN phase—and this suggests that teachers were enacting practices aligned to the TLLM tenet calling for less testing. The macro data also indicated that in the late TSLN phase (2007-2011), Singapore students had teachers who "always or almost always" assessed higher-order cognitive skills like *application* and *understanding*.

The classroom level data, however, differed from the macro level data as assessments were conducted more frequently, at a rate of almost one a fortnight. One reason for this higher frequency of testing was due to a shift in emphasis from a once-off summative assessment towards the practice of documenting students' learning over the school year. Schools had introduced a "continual assessment" policy which computes students' grades over the entire school year into a single score. The introduction of continual assessments was a double-edged

sword. On the one hand, multiple small assessments enabled teachers to have a better sense of what their students knew and were able to do. On the other hand, each assessment was now higher stakes because the marks would be computed as part of the final grade and thus, teachers felt tremendous pressure to test and retest students to ensure that good marks were attained during each assessment. While the use of continual assessments enables documentation of students' learning over the school year, this does not automatically mean that teachers are then able to draw on the assessment data to provide appropriate feedback that can help students clarify their misconceptions and move them to the next step. Indeed, such practices can lead to many "mini-summative assessments" (Klenowski, 2009, p. 264), resulting in a "serial summative data chase" (Lambert & Lines, 2000, p. 134), especially if teachers conduct such tests frequently. In the case of retests that teachers like Maryanne and Miki conducted, students' learning did not improve. Both teachers lamented that their students continued to make the same errors even after they retaught the topic. These teachers' focus on marks and correct answers meant that they did not sufficiently analyze the quality of students' responses as compared to their *more aligned* colleagues, and hence, were unable to identify the most appropriate strategy to help students close the learning gap.

The second *change* pattern was indicated by the attention teachers paid to the assessment of higher-order skills. There was some discrepancy in the pattern because macro and micro data sets examined the assessment of higher-order cognitive skills differently. Specifically, the macro data focused on the frequency of assessing higher-order skills, while the micro data analyzed the extent to which such skills were addressed by the teachers participating in the study. The macro analysis reported the proportion of students whose teachers assessed higher-order skills frequently or infrequently, while the micro analysis examined the proportion of the assessment

task devoted to these skills. The macro data showed that in the late TSLN phase (between 2003 and 2011), more students had teachers who reported that they frequently assessed higher-order skills such as applying knowledge and understanding, and providing explanations and justifications. Assessing students' ability to apply knowledge and demonstrate understanding, and to provide explanations and justifications serve as indicators of the type of higher-order skills envisioned in TSLN. However, at the micro level, based on the rating of assessments using Newmann and Associates' (1996) AIW criteria, this study found that most teachers did not address higher-order cognitive skills extensively in their assessments, even though many of them were aware of the assessment objectives stated in the syllabus document. These data indicated that only a small proportion of the assessment prompts in each task addressed higher-order skills. Thus, while the macro level responses indicated that more students had teachers who frequently assessed higher-order skills, the micro level data showed that only a small proportion of the tasks attended to the assessment of these skills. The micro level data also showed that some teachers were assessing higher-order cognitive skills more than others.

There are several ways to understand the discrepancy between the macro and micro data in the pattern of *change.* The first reason for this discrepancy is methodological. In the TIMSS survey, science comprises biology, physics, chemistry, and earth science. However, because Singapore reports that science in the country is taught as a combined subject, the responses to the Teacher Questionnaire needed to be provided by just the teacher overseeing science. However, in practice, general science and earth science are actually taught by two teachers. Thus, the teachers responding to the Teacher Questionnaire may not be earth science teachers, and thus, their responses may not truly reflect the assessment practices enacted for this subject. A further methodological explanation is that, as with survey items on the frequency of a practice,

respondents may over- or under-estimate the extent of their practices (Cuban, 1984; Mayer, 1999). In addition, when completing self-report surveys, the responses may be selective, lack independent confirmation, and be provided to please specific authorities rather than to offer an accurate picture of classroom practice (Cuban, 1984). Therefore, the macro and micro level data could be jointly used to present a more realistic picture of the classroom practices enacted by teachers participating in this study.

Another difference between the micro and macro data in the assessment of higher-order skills is pedagogical. This study was conducted at the early part of the lower secondary course, when students were aged 13 or 14 years (three teachers submitted work from Secondary 1 students while the rest contributed assignments from Secondary 2 students). At this stage, students were still fresh out from primary school, and were not familiar with the study of geography. Thus, it is possible that, as part of the scaffolding process, the assessments teachers submitted indicated a focus on assessing basic geographical facts instead of paying attention to higher-order skills. In comparison, the macro data were based on responses from teachers teaching Secondary 2 students, and collected at the end of lower secondary education when students would be older, and, thus, might have been presented with more challenging tasks in order to prepare them with the higher-order tasks and prompts needed for the upper secondary level high-stakes examinations (see Chapter 6).

In analyzing the *change* pattern towards the assessment of higher-order skills, the micro data were valuable in examining the extent to which teachers' practices were aligned to the policy intent. The interview data indicated that most teachers' practices—especially those of the *moderately aligned* teachers—reflected the policy to the letter, but not in spirit. For instance, while many of the teachers in this study recognized that being able to elaborate and communicate

ideas were important skills for life outside of school, their emphases on open-ended, constructed-response questions were underpinned by a strategic and pragmatic stance: they had to prepare their students for the high-stakes upper secondary assessments which do not include multiple-choice questions.

Persistence. The third pattern, *persistence*, was evident in the continued use of certain assessment practices over time in spite of the policy encouraging teachers to use more of other practices. As this pattern of assessing discrete facts and knowledge was reflected in both the macro and micro level data, it was, therefore, the most pronounced of the three patterns. In fact, it was the dominant pattern despite the policy intent to develop thinking and reflective students. In terms of the macro data, the pattern of assessing facts and concepts showed a steady increase in frequency in the later TSLN phase from 2003 to 2011, as more and more students had teachers who indicated that they frequently assessed facts and concepts. Likewise, at the micro level, the analyses of teachers' assessments using the AIW criteria (Newmann & Associates, 1996) indicated that teachers focused more on disciplinary knowledge rather than on knowledge construction or knowledge application to new contexts. At the micro level, this pattern of *persistence* was common among all teachers. In comparison, the pattern of *change* was applicable to the *more aligned* teachers and the pattern of *variety* was reflected by the *more* and *moderately aligned* teachers.

In spite of the assessment objectives in the revised syllabuses which stipulate that knowledge should be assessed together with higher-order skills,[32] the micro and macro data indicated that teachers continued to privilege the assessment of content and knowledge rather than addressing application and understanding skills. This emphasis on assessing facts and skills dominated teachers' practices, despite the fact that being able to communicate thoughts and ideas

[32] See Chapter 4 and also Chapter 6, Footnote 2.

and to transfer knowledge to new contexts are some of the traits needed for 21st century living (e.g., Trilling & Fadel, 2009) or for life outside of school (e.g., Newmann, et al., 1995). This finding was surprising, given the fact that that there are real world applications of geography as a subject, that subjects like social studies have higher AIW scores compared to subjects like mathematics (e.g., Gleeson, 2011; Koh & Luke, 2009), and that the syllabus document explicitly mentions the need for students to appreciate the interrelationship between humans and the environment (see CPDD, 2005).

Yet the emphases the teachers in this study paid to the assessing of content are not new. In fact, the analyses of the nature and quality of teachers' assessment reported in this study are similar to the research findings in Chicago (e.g., Bryk, et al., 2000) and Queensland (e.g., Hayes, et al., 2006), in that the quality of assessment tasks teachers presented to their students did not sufficiently challenge their students in the cognitive domains of application, analysis, and evaluation. In fact, many teachers focused overwhelmingly on discrete and isolated facts and knowledge. Furthermore, the micro data showed that, while teachers recognized the value of teaching and assessing higher-order cognitive domains, at the lower secondary level the focus on content and facts persisted because teachers perceived that learning took place in incremental steps, as in the behaviorist learning tradition. Thus, they preferred to first ensure that students had acquired basic content before presenting them with tasks that required the application and transfer of knowledge to new contexts.

While TSLN calls for teachers to use assessments that challenge students intellectually, this does not mean that *all* assessments need to address higher-order skills. In particular, students need accurate and broad-based knowledge in order to handle the complex cognitive tasks and activities that are essential for each discipline (Mullis, et al., 2009). As knowledge and

concepts provide the foundation of learning, Newmann and Associates (1996) included *Disciplined Inquiry* among the AIW criteria. This is also why TLLM does not advocate that teachers abandon their former practices in favor of new ones. However, to be aligned with the TSLN intent, the assessments should not assess knowledge in a discrete and disconnected manner. Teachers should, instead, pay more attention to assessing a range of cognitive domains, and thereby provide more opportunities for students to apply their learning to new contexts and to demonstrate their understanding. However, the analyses of the assessments teachers in this study used found that practices were contrary to the policy intent, even with the explicitly-stated assessment objectives in the syllabus.

Formative assessment

As TSLN and TLLM emphasize learning that should prepare students for and equip them with the skills for life, it was insufficient to confine the study of teachers' classroom assessment to the tasks they presented to their students. Thus, in addition to examining how the teachers elicited student learning through the assessments they presented to their students, this study also examined how teachers analyzed students' responses and made curricular decisions to enhance or improve student learning through the use of formative assessment practices [as reflected in Research Question 2(e)]. For these reasons, this study contributed to the field of classroom assessment by examining the nature of the assessment tasks teachers used, the approaches they adopted to interpreting student work, the way teachers engaged in the formative use of assessment data, and the manner in which teachers employed formative assessment to enhance student learning. In comparison, the large scale studies reviewed in Chapter 2 focused solely on using a rubric to examine the nature of the tasks teachers used to assess student learning.

Both the macro and micro data sets exhibited similar patterns. At the macro level, formative assessment as reflected in the Teacher Questionnaire was used to provide feedback, to diagnose students' learning needs, and to plan future lessons. At the micro level, formative assessment was used for the same reasons. Drawing on the interview comments, the micro data provided additional information on how teachers enacted formative assessment. Specifically, in terms of alignment to the TSLN and TLLM intent, there were variations in the ways in which the teachers enacted formative assessment. The *more aligned* teachers' practices reflected the spirit of the TSLN intent more closely because they enacted formative assessment practices that used open questioning, and focused on metacognition and students' understanding. Such practices resonated with the characteristics of *divergent* assessment (Torrance & Pryor, 2001) in which teachers examine student work for understanding instead of focusing only on correct responses (Torrance & Pryor, 2001). *Divergent* assessment is aligned with social constructivist views of learning (Torrance & Pryor, 2001).

In comparison, the *moderately* and *less aligned* teachers used formative assessment approaches that only reflected the policy intent to the letter. Although they enacted formative assessment, their approach involved the use of "closed or pseudo-open questioning and tasks" and focused on "contrasting errors with correct responses" (Torrance & Pryor, 2001, p. 617). These emphases on rectifying wrong answers and using external rewards to motivate students echo the features of behaviorist learning theories in which teaching proceeds in a linear fashion, and where there is a tight test-teach-test sequence (Torrance & Pryor, 1998, 2001).

There are several explanations for why the teachers adopted convergent practices. First, they experienced time constraints. Several teachers participating in the study said that they were pressed for time, and thus, it was more efficient to give students correct responses or to make

them copy them. Due to a shortage of time, teachers believed that it was simply not possible for them to provide students with exemplars and have them work out the correct answers.

Another reason for convergent practices that emerged from the qualitative data was rooted in teachers' expectations and views of how students learn, and in particular, in the view that students needed to be given correct responses. *Convergent* practices used by the *moderately* and *less aligned* teachers detract from the TSLN and TLLM visions of nurturing active and engaged learners because the teachers continued to direct teaching, learning, and assessment. *Convergent* assessment practices also deviate from the TSLN intent of developing independent learners equipped with lifelong learning skills because when teachers provide the answers and solutions, students become reliant on them and are less driven to engage in any exploration or thinking themselves (Hattie, 2009). Such teaching practices indicate that for these teachers, learning meant *being taught,* in that students' minds, like containers, needed to be filled, and consequently, assessment involved checking if the learning was there (C. Watkins, 2011).

To help students learn, and to develop in them lifelong learning skills, teachers need to help students understand why they have made errors and how they can avoid making the same mistakes (Lambert & Lines, 2000). As such, it is necessary for students to construct meaning by themselves (Lambert & Lines, 2000; Sadler, 1989). To do this, peer and self-assessment are important because students are only able to attain a learning goal if they know and understand the goal, and can determine what they need to do to reach it (Black & Wiliam, 2012a). However, the evidence from the data collected for this study indicates that the *moderately* and *less aligned* teachers devoted their time and effort toward providing students with the correct responses, telling them what was incorrect, and developing strategies to maximize the attainment of marks. Such practices indicate that the type of feedback provided merely focused on the managerial

function of assessment (such as grading and correcting mistakes) at the expense of the learning function (Black, et al., 2003b; Black & Wiliam, 1998a; Gattullo, 2000).

Overall, these practices resonate with the conclusions from Black and Wiliam's (1998a) review of the literature on classroom assessment in the sense that teachers' assessment practices did not address deep learning. They merely attended to the learning of the material and did not concentrate on the meaning or purpose underlying the learning (Marton, et al., 1996). With the exception of the two *more aligned* teachers, teachers used assessments that focused on rote and superficial learning, that emphasized the quantitative aspects of the work, and that valued marks and grades over the quality of responses (Black & Wiliam, 1998a). The evidence from the classroom level data therefore point to the fact that the teachers' formative assessment practices were less directed toward preparing students for the *test of life* than toward subjecting students to a *life of tests*.

Summary

All in all, the combination of the macro and micro data indicates that teachers' assessment practices did not strongly reflect the shifts advocated by the TLLM tenets. Of the three patterns of variety, change and persistence, there was greatest concurrence between the macro and micro data for the pattern of *persistence* in using classroom assessment practices that the policy was calling teachers to use less of. While the macro and micro data exhibited similar purposes in the use of formative assessment, the latter indicated that teachers' practices, such as telling students the correct answers, strongly reflected the pattern of *persistence* as such practices detract from the policy's intent for teachers to do more facilitating and guiding. There was convergence between the macro and micro data for the pattern of *variety* as both data sets indicated that teachers adopted a variety of assessments. There was less convergence for the

pattern of change in that the macro data suggested that teachers were assessing less frequently but the micro data reported an increased rate of testing. The macro and micro data also differed in that the former indicated that there was more frequent assessment of higher-order cognitive skills but the classroom level findings found that over a five-month period, many of the assessments did not address such skills.

The analyses of the three patterns emerging from the macro and micro data suggest that in response to the TSLN vision and TLLM tenets, there was "incremental" as compared to "fundamental" change or reform in teachers' assessment practices (Cuban, 1993, p. 3). Using a metaphor from architecture, Larry Cuban views incremental reform as being akin to remodeling a building or an apartment. This involves making small changes or adaptations to existing structures, which, though sound, require further improvement (Cuban, 1993). Teachers make incremental change because this is one way in which they cope with change as they adopt aspects of the reform that enable them to continue to work smoothly (Tyack & Cuban, 1995). According to Cuban (1984), plausible reasons for persistence in teachers' practices include the way schools and classrooms are organized, the culture of teaching, teachers' ideas about how students develop, and their views about the role of the school and classroom authority.

Cuban's explanations for incremental change resonate with the interview data presented in Chapter 6. For instance, because of large classes and a tight curriculum timeline, the teachers participating in this study said that the most efficient way for them to provide feedback was to "tell' students the correct answers instead of providing opportunities for them to explore on their own and to reflect on their errors. The culture of teaching was such that teachers taught as they were taught, and in this study, they also designed assessment tasks and prompts similar to the ones their teachers had used. In this culture of teaching, teachers view themselves as "the

authority" who dispenses knowledge to students, who in turn have to absorb and digest this information (Cuban, 1993). Additionally, being in an Asian society, the teachers in this study believed that they should wield control and authority in the classroom. Hence, for all the above reasons, the majority of teachers in this study concentrated on assessing discrete content and factual knowledge even though there were numerous opportunities for them to incorporate new contexts into their geography tasks.

Based on Cuban's (1993) definition, the teachers participating in this study exhibited incremental approaches to reform in their assessment practices. This was evident in the use of performance assessments like research projects (e.g., Harry and Totoro), the inclusion of questions that required students to integrate learning across topics (Amanda), and the decision not to assign marks to prompts (e.g., Jiajia and Margaret). These incremental changes that the teachers incorporated into their classroom assessment practices were small and sound, but would require further improvement to be considered fundamental change (Cuban, 1993). In the context of this study, these practices represent incremental change because the majority of teachers did not substantively change the existing classroom culture. For instance, unlike the requirement in the AIW criteria for students to present their work to an "audience beyond the school" (Newmann & Associates, 1996), the teachers in this study continued to be the sole assessors of their students' work. As a group, the teachers also did not involve students in initiating tasks, in co-constructing the assessment criteria, or in self-assessing. Within the group, there was more incremental change among the *more aligned* than among the *moderately aligned* teachers. Totoro and James focused more on the process of learning when they tasked their students to undertake independent research, while *moderately aligned* teachers like Jiajia used assessments that did not assign marks to prompts. Among the three groups of teachers, the *less aligned*

teachers' practices exhibited the smallest extent of incremental change as the assessments they submitted and their interview comments did not reflect the tenets that TLLM was calling teachers to enact more of.

Cuban's concept of *fundamental reform* requires extensive and complete overhaul of a building. In education, this may involve transforming teachers' roles, changing them from being the expert and authority figure in the classroom to being the coach whose role is to guide students in decision-making and to encourage them to learn from one another (Cuban, 1993). In this study, even the *more aligned* teachers did not exhibit fundamental changes to the classroom learning culture as they did not provide time or space for student involvement or decision-making in the classroom. How can fundamental change in which teachers adopt more of the TLLM tenets be brought about? In addition, how can teachers' practices reflect the spirit of the policy intent? I will discuss these issues in the implications section of this chapter.

Student learning

In addition to studying the quality of the assessment tasks that teachers presented to their students, the second phenomenon of interest in this dissertation is student learning, given that TSLN focuses on equipping students with skills that will enable them to learn continuously throughout their lives (C. T. Goh, 1997). This dissertation used different indicators to analyze the quality and quantity of learning. First, Singapore students' achievement scores in five TIMSS cycles were used as indicators of the quantity and quality of learning over time. Student learning, as presented in Chapter 4, was reflected in the overall science and cognitive domain scores. Second, the AIW student work criteria (Newmann & Associates, 1996; Newmann, et al., 1995) were used as indicators of the quantitative and qualitative learning that students are able to demonstrate, as these are similar to the skills emphasized in the TIMSS cognitive domains and in

the assessment objectives of Singapore's lower secondary geography syllabus. The latter were designed to be aligned with the TSLN intent.

As suggested in Chapter 4, at the macro level, Singapore students' sustained performance in five TIMSS cycles indicates the quality of learning, and shows that students can handle assessment questions for which they have not been specifically prepared. In addition, another indicator of the quality of student learning is that Singapore students performed better in the *reasoning* domain than in the *knowing* domain in TIMSS 2011 (Martin, et al., 2012). This achievement may be attributed to Singapore's curricular emphases on inquiry-based teaching and learning in schools over the years (Ministry of Education, 2012). Additionally, the evidence from the Test-Curriculum Matching Analysis (TCMA) data further suggests that Singapore students have demonstrated that they are able to apply their learning in the biology, chemistry, earth science, and physics components in TIMSS Science, despite the fact that for each TIMSS cycle, only about three-quarters of the score points in these international test items were relevant to the Singapore Secondary 2 curriculum (Table 4.13). Therefore, the evidence drawn from the macro level data suggests that Singapore students are able to apply concepts learned to new contexts.

The findings from the micro level data are consistent with those from the macro data, in that students are able to apply their learning to higher-order tasks when given the opportunity to do so. The statistical analysis presented in Chapter 5 indicates that there was a direct relationship between the quality of teacher assessment and student work. Specifically, when the teachers designed assessments that demanded higher-order skills, correspondingly, the responses students produced exhibited higher-order skills. Thus, when the *more aligned* teachers and some of the *moderately aligned* teachers presented their students with tasks that required the synthesis

and analyses of data, or the application of content to the real world, their students produced responses that indicated their ability to demonstrate higher-order skills. However, in the course of the research, this study found that the majority of teachers did not provide their students with such challenging tasks. In particular, since many of the assessments only required students to produce short responses or to reproduce factual content, there was little opportunity for students to demonstrate their ability to address higher-order tasks as required in the AIW rubric.

The micro findings support the research examining the relationship between teacher assessment and student work (e.g., Clare & Aschbacher, 2001; Gleeson, 2011; Ladwig, et al., 2007), in that there is a direct relationship between the quality of assessment tasks presented to students and the quality of student responses. This means that the nature and quality of assessment tasks teachers use impose a ceiling effect on the quality of work that students produce (Koh & Luke, 2009). Therefore student learning and student achievement are related to the quality of assessments that their teachers use (Gleeson, 2011). When an assessment makes low demands on authentic work, students will most likely get a low score because they will have no chance to demonstrate their proficiency in any of the AIW criteria (Newmann, Lopez, et al., 1998). Conversely, when teachers have high expectations and create prompts that require students to demonstrate authentic work, there will be opportunities for students to show what they can do, and hence, the quality of work that they produce will increase (Koh & Luke, 2009; Newmann, Lopez, et al., 1998). In light of this, teachers should consider assigning their students more challenging assessments which are developmentally and instructionally scaffolded.

At the macro level, Singapore students' performance in the content and cognitive domains lends support to the assertion that when provided with challenging tasks that are appropriately scaffolded, students are able to apply their prior knowledge and provide high

quality responses. This is evidenced by their sustained strong performance in the cognitive and content domains in each TIMSS cycle. Yet, the micro data indicated that the majority of the teachers in this study did not assign challenging tasks to their lower secondary students. The teachers did not do so because they viewed lower secondary students as lacking in content and background knowledge. As a result, the micro data analyses indicated that the students did not produce work exhibiting the types of higher-order skills required in the AIW-derived rubric.

Based on the findings, there appears to be a discrepancy between the macro and micro data on student learning, with the former indicating that Singapore students had performed well in the content and cognitive domains, and the latter providing evidence that student work was not exhibiting higher-order skills. There are three possible reasons for this discrepancy.

The first reason is due to the instrument used to analyze the quality of student work. The AIW-derived rubric that examined student work at the micro level had three standards, *analysis*, *disciplinary concepts*, and *elaborated written communication*. While each standard carried a rating of 4, the rubric seemed to privilege work that involved extensive written communication. By extension, tasks in which students were engaged in extended writing such as projects would receive higher AIW scores. However, during the period this study was conducted, several teachers were teaching Map Reading. The assessments for this topic only required students to demonstrate their ability at the *analysis* standard as they had to interpret and organize geographical data from the maps. There was little extended writing required as most of the prompts only required short, one word responses. As a result, student work for Map Reading tasks did not receive high AIW scores. On their own, the pieces of student work received credible marks and grades from their teachers, as reflected in the teachers' reports of the percentage of A-grades which their students scored. Thus, based on the actual marks, students

showed that they were able to handle higher-order skills like *analysis* and *disciplinary concepts*. However, the completed student work was unable to receive high AIW scores because there were insufficient opportunities for students to engage in extended written work. As a result, the student work scores reflected lower quality student work. Comparatively, assessments that were based on human geography topics such as pollution and population required students to respond to issues-based prompts using extended writing. As such, students' responses to such topics received higher scores for the *elaborated written communication* criterion, thereby increasing the overall AIW score. In light of this, future research could analyze and compare the AIW scores for student work on human and physical geography to ascertain if there is indeed a difference due to the nature of the curricular content.

In addition, the AIW-derived rubric was not applied to student responses to the TIMSS assessment tasks. Instead, student achievement in the TIMSS assessments was based on responses to the prompts presented, not on this external rubric. The TIMSS assessments are comprised of a combination of multiple-choice and constructed-response prompts. If students' responses to the TIMSS prompts had been rated against the AIW-derived rubric, it is possible that student responses to the TIMSS multiple-choice prompts would be rated high on the *analysis* and *disciplinary concepts* standards, but low on the *elaborated written communication* standard. This would then lower the overall quality of student work. To this end, future research could perhaps use the AIW-derived rubric to examine student responses to the assessment blocks that TIMSS releases at each cycle[33] in conjunction with the assessments their teachers present. This would ensure that a common yardstick is used to examine the quality of student work based on an international assessment as well as on the tasks that teachers use.

[33] Released items are provided at each TIMSS cycle together with the scoring guide for the constructed-response items. For an example, see http://timss.bc.edu/timss2011/international-released-items.html

Another reason for the discrepancy between the macro and the micro data can be attributed to the nature of large-scale standardized tests. While it is true that school systems are unable to prepare students for international benchmarking tests like TIMSS and PISA, these tests are limited in the extent to which they are able to assess a range of student abilities (Atkin & Black, 1997). Scholars have criticized the nature of benchmarking studies because they are comprised of multiple-choice questions and constructed-response questions. In comparison to the demands of the *elaborated written communication* standard in the AIW-derived rubric which requires that students elaborate, take alternative points of view, and provide a coherent argument, the constructed-responses from these large scale tests only require short, "closely defined" answers (Atkin & Black, 1997, p. 25). Given the nature of these large-scale tests, scholars (e.g., Hogan, 2014) who are familiar with the Asian systems point out that transmission teaching methods, such as those used by the teachers in this study, are able to generate outstanding performance on international tests. Furthermore, in Confucian Heritage Cultures like Singapore there is a culture of equipping students with skills to navigate standardized assessments (Biggs, 1996b). Compared to performance tasks like research projects which require students to engage in data collection, interpretation, and synthesis and organization of information, large-scale assessments do not measure process outcomes (Atkin & Black, 1997). Possibly the only international study that examined such skills was the 1995 TIMSS performance Assessment which was not conducted in subsequent cycles.

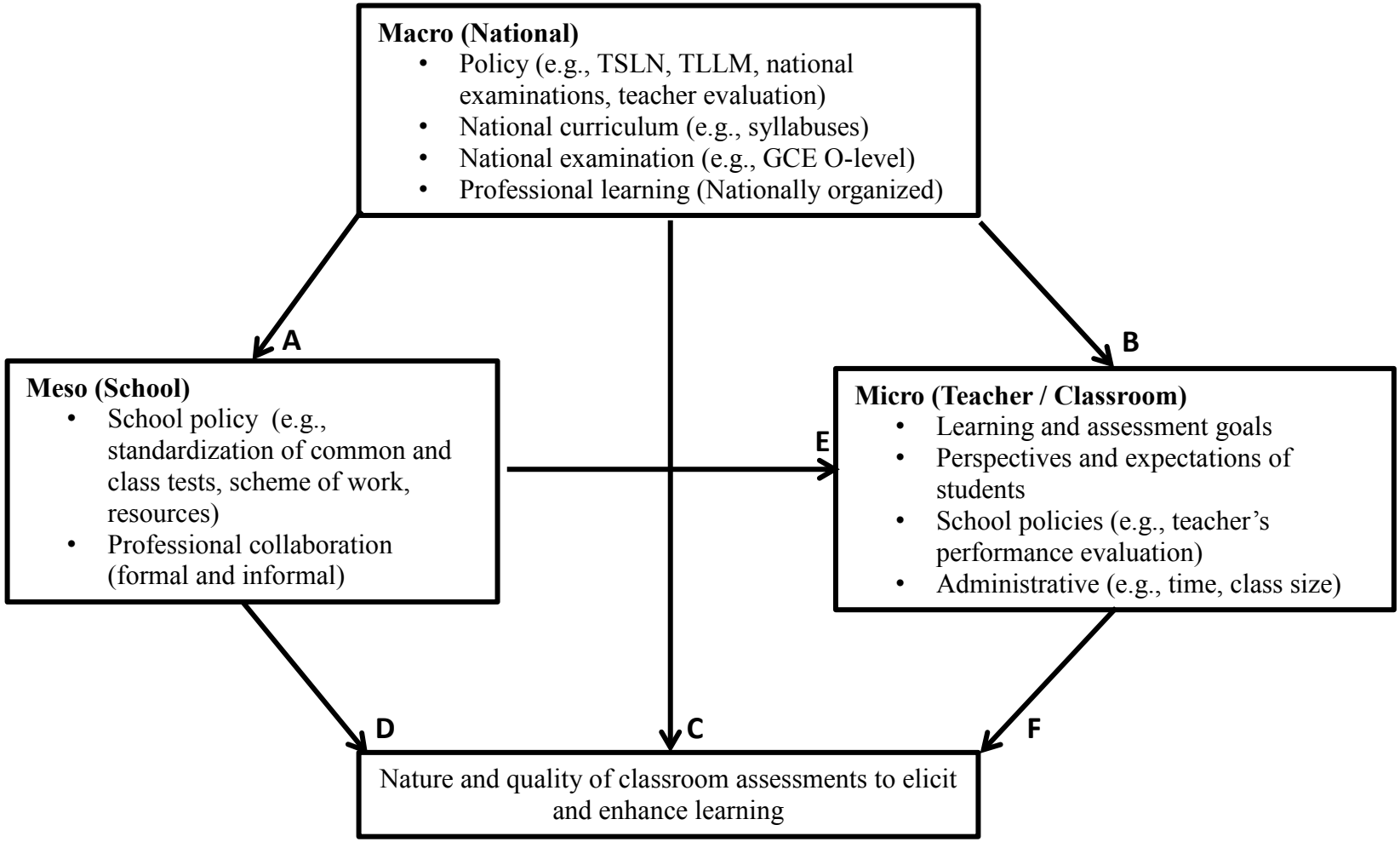
Third, it is possible that the micro data scores for student work were lower because of the types of assessments the teachers submitted. While the "kit" outlining the research procedures, which was given to the teachers, instructed them to submit a 'culminating' assessment, many teachers did not adhere to this requirement. A 'culminating' task was one that required students

to demonstrate geographical understanding at a higher level, or integrates different aspects of one or more topics (see Annex B). Due to the duration of the study, and to teachers' department schedules, not all the assignments submitted were culminating ones, as defined above. Several teachers submitted worksheets that pertained to a particular section of the topic that they were teaching. As a result, student work completed in response to these assessments scored low as these mid-topic tasks focused only on discrete content knowledge. Future research would remedy this by having a longer data collection period, which would enable teachers to submit end-of-unit assessments that make higher cognitive and content demands on students.

Framework of factors influencing assessment practices

As this chapter presents the meta-inference of the macro and micro data, Figure 7.1 integrates the findings on the influences on the teachers' classroom assessment practices from Chapters 4, 5 and 6. Broadly, the nature and quality of classroom assessments are influenced by three levels, namely, national policy (*macro*), school (*meso*), and teacher (*micro*). The extent to which teachers' classroom assessments reflect the policy intent depends on the interaction of the macro, meso, and micro levels.

Figure 7.1
Macro, meso and micro factors influencing lower secondary classroom assessment



At the **macro** level, overarching educational policies such as the TSLN vision set the broad direction to drive reforms. Such policies are publically communicated to schools (Arrow A) and teachers (Arrow B). In Singapore, another significant macro element that is relevant to this study is the national curriculum which is manifested in the subject syllabuses. The national curriculum and examination policies articulate the knowledge, skills and nature of assessment at the key and non-key stage levels (Arrow C). In recent years, this includes the introduction of project work at the GCE A-level to signal the importance of assessing research skills and oral presentation (Y. K. Tan, et al., 2008) as well as the stipulation of assessment objectives to guide teachers on the assessment of different cognitive domains (see CPDD, 2005). Aligned to the spirit and intent of TSLN which aimed for schools to be milieux of innovation, there are no prescribed procedures for translating the TLLM tenets in school. Therefore, schools have the autonomy to customize the syllabuses according to their student profiles. At non-key stage levels such as lower secondary, there is flexibility for schools to determine teaching approaches as well as the nature of school and classroom assessment. On the one hand, this flexibility has accorded schools more autonomy to plan instructional activities. On the other hand, the autonomy given to schools has resulted in a great diversity of practices, with some being more aligned to the spirit of the policy than others. The discretion afforded to assessment practices that move towards TLLM is especially important given that considerably less discretion is afforded to other aspects of assessment policy, especially those involving secondary school examinations.

A range of macro policies drive practices in schools and in classrooms. Within the overarching TSLN vision, TLLM was a pedagogical policy, and hence, it focused on the *what*, *why*, and *how* of teaching. TLLM's goal of preparing students for the *test of life* was at times hindered by other policies, such as the national examination and the performance-related teacher

evaluation policies, both of which affect decision-making in schools and in classrooms. Together, these other policies have created tensions and contradictions at the macro level which have ultimately affected the extent to which schools and teachers enacted the spirit of TLLM. In its call for teachers to use more of some approaches and less of others, TLLM appears to be optional while the national examinations and teacher evaluation policies are mandatory and carry with them high stakes. To this end, at the macro level, TLLM's influence is being undermined by these other policies, as they manifest themselves at the *meso* level of schools.

At the **meso** level, school policies directly and indirectly influence teachers' assessment practices. Policies that directly affect assessment practices include school and departmental decisions on the number, frequency, and nature of assessments to be implemented, as well as the way assessment information is to be used (Arrow D). In most instances, these policies are purely administrative and they have resulted in teachers adopting standardized assessment procedures. In some schools, common tests and classroom assessments were administered as mini-standardized assessments. The standardization of assessment processes and the use of standardized common tests in some schools meant that teachers had less flexibility to tailor assessment practices to students' needs and to their learning goals. The clockwork precision with which departments planned teaching and the implementation of worksheets, tests and examinations, made it impossible for some teachers to provide more individualized assessments and feedback for their lower secondary students. While some of these school policies are designed to make testing more efficient, rigorous and fair across all classes, they have led teachers to rush through their teaching to keep abreast with the schedule, and they have therefore been less able to customize assessment and feedback to their students' needs. Some of these centralized school procedures detract from the TSLN spirit of encouraging more flexibility and

customization of teaching to students' needs. These school processes take direction from efficiency and productivity and resonate with the efficiency era in education—a different curricular paradigm from that positioned in the TLLM tenets.

In addition, school policies such as the use of students' scores to evaluate teachers indirectly led to their enacting assessment practices that detracted from the policy intent (Arrow E). Some schools persisted in using student performance as the sole indicator of teacher effectiveness, despite the fact that the teachers were teaching a non-key stage level, and despite the fact that teacher evaluation in Singapore is comprised of a range of other indicators (OECD, 2011). In schools that adopted such evaluation policies, teachers responded—pragmatically—by adhering very closely to the syllabus, by adopting drill and practice to ensure that students were able to swiftly and confidently reproduce formulaic responses, and by designing assessments that mimicked the examination format so that students became comfortable and familiar with test conditions.

On the other hand, school policy and school culture can also indirectly impact teachers' assessment practices in a positive way. Some schools had generous resource policies in terms of funding and staff assignment which teachers could draw on to support the conduct of alternative assessments. In other schools, the culture encouraged formal and informal professional collaboration and learning. This enabled teachers like Harry to work with colleagues to design creative and innovative assessments that involved independent student research and assessed interdisciplinary learning. In these instances, school policy was more aligned with the macro policy of TLLM in encouraging innovation and change at the grassroots level with the teachers working together to construct assessments that addressed higher-order skills.

At the **micro** level, classroom assessments were influenced by individual teachers' expectations and views of their students, and their curricular goals (Arrow F), which may or may not take reference from the macro policy. When these views of students and curricular goals were aligned with the policy and its constructivist underpinnings as envisaged in Shepard's (2000) "emergent" paradigm or Torrance and Pryor's (2001) *divergent* assessment, the teachers used a greater variety of assessment modes and formats, addressed a wider range of cognitive skills, and presented their students with more challenging tasks. Their formative assessment practices required students to make sense of their misconceptions and errors in order to arrive at a deeper level of understanding. In general, however, apart from the two *more aligned* teachers, and occasionally, Margaret and Jiajia, the other teachers did not guide students to make sense of the misconceptions by themselves, and to build their learning from these.

Most of the teachers involved in the qualitative component of this study had views of learning, of assessment and of their students that were aligned with the "dominant 20th century paradigm" (Shepard, 2000). This is characterized by a view that knowledge is sequential and hierarchical, by the use of positive reinforcement (e.g., praise, rewards) to motivate students, and by the adoption of test-teach-test strategies to ensure learning (Shepard, 2000). As a result, these teachers valued objective tests (Shepard, 2000) over qualitative assessment, and they employed close questioning, and focused on comparing errors with model answers (Torrance & Pryor, 2001). The findings from the micro level influences on teachers' assessment practices indicate that as a group, the majority of teachers in the qualitative part of the study held views of teaching, learning, and assessment that detracted from the policy intent.

Overall, the interrelationships among the three macro, meso, and micro levels influence the extent to which teachers' assessments were aligned to the policy intent, in terms of whether

they focused on learning as envisioned in the policy, or as marks and grades. Teachers whose assessments were *more aligned* to the policy goals spoke frequently of, and had practices which resonated with the broad policy parameters in that they believed all lower secondary students can learn and can handle higher-order skills. These teachers also considered presenting their students with more challenging tasks. In addition, these teachers worked in schools that had policies—administrative, curricular, or resourcing—that had been developed and implemented in alignment with the macro policy. At the personal level, these teachers had learning goals that matched the macro policy in that they facilitated and guided their students towards deep learning and understanding. The interaction of the macro, meso, and micro levels resulted in teachers whose assessment practices focused on learning rather than achievement. These teachers focused on the quality of student responses instead of emphasizing the percentage of grades and marks. In asking their students to reflect on their work and to make sense of their own errors, these teachers' classroom assessment practices focused on *learning is the process of sense-making* (C. Watkins, 2011). Such practices resonate with newer paradigms of learning (Shepard, 2000; C. Watkins, 2011), which are implicit in TSLN and TLLM.

Conversely, at the other end of the continuum, the *less aligned* teachers had views which reflected the aspects that the policy was encouraging teachers to use less of. Specifically, their strategies focused on "telling" students correct answers, and they emphasized grades and marks. In addition, they did not pay attention to the guides and directions provided in the curricular documents. Furthermore, these teachers also did not heed the assessment objectives stipulated in the national syllabuses. For these teachers, assessment and other policies in their schools were not developed or implemented to be in alignment to TSLN and TLLM. In particular, their schools adopted teacher evaluation policies that emphasized marks and grades that students

attained. The interaction of the macro-meso-micro levels meant that these teachers addressed achievement and performance rather than learning. These teachers' assessments indicated that they paid attention to the conception that *learning is being taught* (C. Watkins, 2011), in which they expected their students to be containers waiting to be filled and to absorb knowledge and facts passively. In short, these classroom assessment practices diverged from the policy intent as they resulted in students being dependent on their teachers for answers and solutions instead of developing dispositions and attitudes to be independent and lifelong learners. When students have not been active in constructing the learning themselves, they will be unable to retain the learning for a longer period of time, as they have not engaged in "deep learning" (Biggs, 1996a; Lambert & Lines, 2000). This explains why students repeatedly made the same mistakes, much to their teachers' puzzlement and consternation.

Located between the *more* and *less aligned* teachers on the continuum, the *moderately aligned* teachers' experiences were mixed. At the micro level, these teachers' views on teaching, learning, and assessment mostly reflected the TSLN intent, but their practices were in alignment only to the letter. Thus, while this group of teachers was cognizant of, and spoke of the policy intent, teachers' practices were not reflective of the spirit of the policy. Overall, their view of learning, like the *less aligned* teachers, was that of *learning is being taught* (C. Watkins, 2011), and thus, they too, focused on test preparation, ensuring that students had the correct answers. At the meso level, these teachers worked in schools that had in recent years adopted procedures to turn classroom assessments into mini-standardized assessments, thereby reducing the extent to which teachers could tailor teaching and learning to their students' needs.

Figure 7.1 provides a unified structure to understand the teachers' assessment practices. The analyses of the responses to the Teacher Questionnaire survey provide the national picture of

classroom assessment practices, and these may be understood from the factors affecting the macro level. The findings from the micro data serve two purposes. First, they present a glimpse of eight teachers' classroom assessment practices. Second, the micro data provide explanations for these teachers' practices which point to influences at the policy, school, and classroom levels. The macro, meso, and micro levels could be examined individually or in concert. When analyzed individually, the teachers' comments throw up tensions within each level which hinder the assessment practices from being aligned to the policy intent, either in spirit or to the letter. For instance, the qualitative analysis of the micro data indicate that the high-stakes national examination and teacher evaluation polices drive teachers towards assessment practices that subject students to a life of tests and retests. Thus, greater coherence within the different macro level policies must be created to enable teachers to enact assessment practices that better reflect the spirit of TLLM.

When combined together, the interaction of the three levels presents a paradox of central loosening and local tightening. On the one hand, macro policy encouraged school-level innovation so as to better respond to students' needs. On the other hand, schools adopted more standardized practices which affected the extent of flexibility and customization that teachers could provide for students to support learning. As a result of this tension, the data indicated that greater centralization at the school level resulted in the majority of classroom assessment practices deviating from the policy intent. The implications, arising from the interrelationships among the macro, meso, and micro levels, point to the need to find ways to align the meso and micro levels with the macro policy. For example, how can teachers' adopt views and enact practices aligned to the policy intent? How can policy makers help schools and teachers understand the philosophy of the policy?

Conclusion of findings

In concert, the macro and micro data indicate a policy-practice gap. This is because both data sets showed teachers' *persistence* in the assessment of facts and content. While both the macro and micro data indicated that there was *variety* in teachers' assessment practices, only the macro data indicated that there was *change* in the assessment practices.

The combination of the macro and micro data has provided a deeper picture of the patterns of classroom assessment compared to each data source on its own. While the macro data captured patterns at the national level, the micro data presented the actual classroom practices as well as provided insight into the rationale and purposes behind the practice, and enabled the examination of the extent to which the practices reflected the policy intent. In concert, the macro and micro data suggest that there are aspects of classroom practices that may be almost immutable (Cuban, 1984, 1993). While TSLN and TLLM represented new policy approaches, teachers' views on teaching, learning, and assessment had yet to be aligned with the reforms.

This study of teachers' assessment practices has contributed to the field of classroom assessment in several ways. First, the studies on teacher assessment and student work reviewed in Chapter 2 were based on school and district programs rather than on a reform that is part of a centralized national curriculum. This study has extended beyond the extant research by examining teachers' assessment practices in the context of an educational policy and national curriculum that have, for over ten years, been articulating and further defining its principles, intent and vision to give more guidance to teachers.

Second, while the research focused on the nature of the tasks presented to students to determine the quality of student learning, this study went further to examine the way teachers

used formative assessment to enhance student learning after reviewing students' work. Typically most studies focus solely on the assessments teachers use or the way formative assessment is enacted. However, this dissertation combined both the presentation of tasks and the provision of feedback to examine the way teachers elicited and enhanced student learning. The approach taken in this dissertation has value in presenting classroom assessment as the combined processes of planning the tasks assigned to students and interpreting assessment information to support student learning. These two processes need to be aligned in order to support and enhance student learning in the way intended by the policy.

Third, this dissertation examined classroom assessment in geography, a subject that is not extensively studied in the research. The analyses of geography assessments provides an indicator of whether teachers were requiring students to apply knowledge and concepts learned to new contexts, given that geography is the study of places and environments. While the research has reported that subjects like social studies had higher AIW scores, meaning that there were more opportunities for teachers to address AIW skills related to *Disciplined Inquiry*, *Construction of Knowledge*, and *Value Beyond School* in their assessments, this study found that teachers continued to address *Disciplined Inquiry* only and provided little opportunity for students to engage in higher-order geography skills such as explaining the "relationship between the earth and the life that exists upon it" (J. R. Smith, 1928, p. 10). Moving ahead, one must ask how teachers can be guided to construct assessment tasks that pay more attention to the *Construction of Knowledge* and *Value Beyond School* criteria.

Singapore's central policy articulated the vision for and direction of change. Compared to earlier reforms from the efficiency-driven era which adopted prescribed textbooks and teaching strategies (OECD, 2011), the implementation of TSLN and TLLM was different

because only broad directions were articulated for classroom practices. Specifically, because the period of "large fixes" was over (Shanmugaratnam, 2005b), the policy allowed for flexible implementation with schools having the autonomy to interpret and implement the broad guiding parameters. However, the greater autonomy accorded to schools has led to a variety of classroom practices, with some teachers enacting practices that are more aligned to the policy vision and rhetoric, and other teachers' adopting practices that detract from the reforms to a large extent. Therefore, one implication is for policy makers to review the way a policy vision is implemented. For instance, what would be an appropriate balance between central control and local autonomy? And more specifically, what would coherent and consistent policies that allow for qualitative and quantitative assessing look like?

Overall, the macro and micro data indicate that Singapore teachers adopted a hybrid of assessment approaches, a common practice in educational reform. Typically in the light of change, teachers enact incremental reform as this does not require them to undertake massive and extensive changes to their existing practices (Cuban, 1993). The evidence from the macro survey and classroom data collected for this study indicates that teachers' assessment practices reflect incremental change because they incorporated alternative assessment formats to gather evidence of student learning. The classroom data further indicated that other than these small adaptations, there were no major differences in the way assessments were conducted: teachers still controlled and directed the assessments, students continued to complete individual pieces of work under timed conditions and played the role of passive participants, and teachers and students still viewed assessment quantitatively rather than qualitatively. Conversely, examples of fundamental change—by Cuban's (1993) definition—would involve students jointly constructing the assessment with the teacher, being given some opportunities at decision-making,

or working in groups rather than as individuals on an assessment task. Another example of fundamental change requires students to present their work to an audience other than the geography teacher, as Newmann and Associates (1996) envisaged in the AIW criteria. The discussion under, *Implications*, will present suggestions for bringing about fundamental change in teachers' classroom assessment practices.

Limitations of the study

The use of multiple data sources in this dissertation over several time points enabled triangulation of the findings. Despite the care and consideration paid to the design of the study, there are several limitations which are addressed in this section.

Limitations in macro (survey) data

First, there are limitations concerning the use of self-report survey data (Mayer, 1999). In particular, teachers responding to the TIMSS survey data may over- or under-estimate the frequency of their practices and assessment decisions for various social, cultural, or professional reasons. While cognizant of this, the aggregate responses of the teachers teaching sampled students randomly selected by the TIMSS sampling algorithm minimize this error. In addition, as the survey is not high-stakes in determining promotion or compensation, survey responses are likely to be more accurately presented. Nevertheless, it is possible that for various reasons, teachers may consciously or unconsciously over- or under-estimate the extent and frequency of their practices. Future research could apply the same survey but triangulate teachers' responses with interviews and classroom observations.

Although teachers' responses to the "assessment" section of the TIMSS Teacher Questionnaire were analyzed, over the five cycles, there were no common questions which spanned the five cycles. One reason for this is that the Teacher Questionnaire, and the other

contextual questionnaires, is revised for each cycle, based on consultation with the international representatives. With the exception of the frequency of testing and cognitive domains categories that covered three cycles (2003, 2007 and 2011), the other survey items spanned just two cycles. Future research could analyze teachers' classroom assessment practices over time using just one version of the survey in order to more accurately study changes in practices over time.

Limitations in micro (classroom) data

A study of changing assessment practices should, ideally, be conducted together with classroom observations. For this study, classroom observations would provide more in-depth information on the way teachers enacted formative assessment. However, in this dissertation, due to time and resource constraints, the study of teachers' assessment practices did not involve classroom or school visits, which would have provided more evidence for the analyses of teacher assessment and student work. Furthermore, classroom observations of teachers' formative feedback conversations with students would strengthen the examination of how teachers seek to help their students *learn more*. Thus one weakness of the inferences made from this study is the absence of the classroom-based evidence. Future research involving classroom observations as well as interviews with students would strengthen or deepen the findings.

Second, because of logistical (e.g., storage) and practical (e.g., not being able to be present in the classroom) limitations, only written versions of student work that were in A4 format are analyzed. In the spirit of authentic assessment, students can demonstrate their learning through other forms of communication (e.g., oral, performance, and electronic) which were not collected for this study. This is one limitation in the data collection approach that could be remedied in future research. In the 21st century, there are many ways to communicate ideas

and learning. The research field would be enriched with analyses of student work that are completed using a range of media platforms.

Third, due to time, resource and practical constraints, this study focused only on geography. In the extant research, some studies have focused solely on mathematics (e.g., Ohlsen, 2007), while others (McMillan, 2001; Stiggins & Bridgeford, 1985; Zhang & Burry-Stock, 2003a) analyzed teachers' classroom assessment practices for different subjects. Other research has also compared variations in the assessment practices based on grade levels (e.g., Trepanier-Street, McNair, & Donegan, 2001), years of teaching (e.g., Suah & Ong, 2012), and subject area (e.g., McMillan, 2001) which could be further explored with more teachers participating in the study. This would enable a comparison across different subjects which would enable the examination of whether the views about teaching, learning, and assessment gleaned from this dissertation are peculiar to geography or are applicable to other subjects. The decision to focus on geography in this study might have limited the examination of teachers' classroom assessment because this is a new subject for students at the secondary level in Singapore schools. As students did not study this subject at the primary level, it is possible that teachers adopted a step-by-step approach to teaching, learning, and assessing this subject at the lower secondary level, beginning with attention paid to facts and content. This could have had an impact on the nature and quality of the teachers' classroom assessment practices, as well as on the quality of student work.

Fourth, the AIW were intended to be broad enough to be applied to different subjects and grade levels (Newmann & Associates, 1996). While the QSRLS and CRPP-CPP studies reviewed in Chapter 2 created and adapted the AIW, these instruments are also generic. However, given that each discipline has its own distinguishing characteristics, it would be useful to create

and validate an instrument that is relevant to geography for rating and examining the quality of assessments for this discipline. An instrument specific to analyzing geography assessments would be particularly useful, because as a discipline, this subject unites the social and natural sciences, which could involve additional cognitive demands.

Fifth, Singapore's lower secondary program runs for two academic years, spreading over four semesters. However, this study only documented teachers' assessment practices over five months. To this end, more comprehensive data and analysis could be obtained by studying teachers' assessment practices over the two-year lower secondary period. This would enable an examination of the nature and quality of physical and human geography assessments, given that the syllabus includes both components over a two-year period.

Sixth, Newmann and Associates did not develop criteria to establish students' views of the tasks they completed for the student learning aspect of the AIW rubric. The other studies (e.g., the QSRLS) also did not examine the value and meaningfulness of the assessments from the students' perspective. This is one gap in the existing field, and also in this study, that warrants further examination by researchers.

One other challenge in this mixed methods study was the sample size. Given the time and resource constraints of this study, having a small group of participants was one way to examine teachers' classroom assessments deeply. However, this small sample size was a limitation for the quantitative analyses conducted in Chapter 5. For instance, it was not possible to conduct statistical analyses such as repeated measures analysis of variance (ANOVA) to examine if there were differences in teachers' classroom assessments over the three time points of the study. An *a priori* power analysis indicated that the sample size was not sufficiently large

enough to provide the power needed to rule out statistical error. Thus, further research could be conducted with a larger sample size to include additional data points.

Implications

The broad findings from this study are that after a sustained 15-year period, the TSLN rhetoric and vision largely resulted only in incremental changes in most teachers' classroom assessment practices. While teachers concur with and value the importance of preparing students for the *test of life*, their assessment practices mostly converged towards preparing students for a *life of tests*, thereby departing from the policy intent. This is evidenced by frequent testing and retesting, the focus on achievement and marks attained, and the nature of the assessments used. Based on the macro and micro data, the findings of this study have implications for policy, for research and teacher educators, and for practice in schools.

The discussion and exploration of the implications are directed toward moving teachers' practices from incremental toward fundamental change. By Cuban's (1993) terms, examples of the latter include transforming teachers' roles and changing teaching cultures. The discussion in this section also aims to change the one-directional arrows in Figure 7.1 to two-directional arrows, such that the interactions among the macro, meso, and micro levels are reflexive and iterative. This will enable change to be more of a ground-up initiative and less of a top-down directive.

Implications for policy

The findings from this study present somewhat contradictory information to policy makers. On the one hand, the macro data usher in encouraging news: at the national scale, and over three to four time periods, the survey data indicated that students were being taught by teachers whose assessment patterns exhibited the combined patterns of *change* and *variety*.

These two patterns of *change* and *variety* provide evidence of some shifts in practices toward the policy intent. On the other hand, the macro and micro data raise issues that require policy makers' attention, in that teachers persist in using traditional assessment practices, even though they are cognizant of the policy vision and objectives. Third, the micro data points to the diversity of practices among teachers and schools which would require more support from policy makers in order to realize the TSLN goals. Finally, the micro data also reveal tensions and contradictions within the macro policy areas of accountability, teaching, and evaluation that need reviewing and resolving. Otherwise, goals and objectives that conflict and compete with the policy intent will result in further divergence from the policy intent.

Establish coherence among different policy elements. The first necessary aspect for policy makers' attention is to address contradictions at the macro level. The teachers' comments about teacher evaluation and the need to ensure students do well in summative examinations at the school and end-of-key-stage levels brought to light competing policies at the macro level, which compelled them to respond pragmatically. In spite of TLLM's call for *more* qualitative assessing, the teacher evaluation and national examinations policies still hinge heavily on quantitative measures such as the MGS. In this sense, while TLLM, as a pedagogical policy, is leaning toward constructivist approaches, the other policies appear to continue residing in behaviorist traditions. Since TLLM only calls on teachers to enact more of some strategies and less of others, it is therefore not surprising that pragmatic teachers persist in using tried-and-tested approaches to attain the dual goals of getting a good evaluation as well as obtaining credible student examination scores.

To this end, if policy makers are committed to realizing TLLM's intent of preparing students for the *test of life*, then coordination and revisions are required at the macro level to deal

with these competing policies. In calling for teachers to use more formative and qualitative assessing, TLLM is encouraging teachers to adopt more alternative assessments and to reduce the emphasis on summative assessments. While this call has high symbolic importance as a stated desire and direction, it is—at the end of the day—voluntary and discretionary in terms of how teachers and schools interpret and implement it. While the nature of the TSLN and TLLM policies suggest there are change and variety at the policy level, the teachers' interview comments indicate that there is also persistence from the efficiency era in the nation's assessment strategies. This is especially evident in the continuation of high-stakes national examination systems, and the continued use of teacher evaluation tools that are linked to performance results. In comparison to the TLLM tenets, the persistence of these assessment orientations of the efficiency era is not voluntary or discretionary, nor does it fall within the realm of school autonomy. The persistence of these traditional assessments is high stakes and mandatory. In other words, at the policy level, not just at the classroom level, there needs to be not only more of the alternative assessments but also less of the traditional assessments, with a distinct and determined effort to reduce the impact of traditional assessments on teachers within the existing educational system.

Therefore, instead of a once-off pen-and-paper examination, there could be a school-based component that allows teachers to design alternative tasks that encourage process skills rather than just "examination learning" (Hogan, 2014). Including aschool-based component is akin to Hong Kong's introduction of School-based Assessment (SBA) in an effort to push for a more balanced approach to a highly pressurized education system (Berry, 2011). The Hong Kong government's goal in adopting SBA in 2001 was to reduce emphasis on summative

assessment, encourage the use of quality feedback in formative assessment, and involve students as active partners in the assessment process (Berry, 2011).

Pushing for such policy changes to examinations and assessment requires policy makers to make bold decisions to slaughter or " tweak sacred cows" (Hogan, 2014) as well as to manage public opinion and expectations while searching for alternatives to reduce emphasis on an examination-driven culture. Hong Kong's experience in assessment change provides testimony to the scale and difficulty of this task as more than a decade later, this jurisdiction continues to struggle to realize its goals for examination reforms.

Singapore's educational reforms, like Hong Kong's, have focused on reducing examination pressure. MOE has, in recent years, signaled its commitment to reduce the emphasis on performance indicators. For instance, it abolished the banding of schools by academic results in 2012, and announced that there will be new approaches to recognize good schools and their best practices (Heng, 2012). In making this decision, MOE has strongly signaled its commitment to achieving the TSLN vision of preparing students for life outside of school. Similar approaches to review the existing examination and teacher evaluation policies with the pedagogical directions envisaged in TLLM will indeed be tentative steps in the journey toward bringing about fundamental change in classroom assessment practices.

In terms of aligning the policies, the second area that policy makers need to address is to monitor more closely the policy implementation at the meso level and how it is communicated to the micro level. TSLN envisions schools becoming the seedbeds of ideas and innovations so that change no longer emanates top-down from policy makers (C. T. Goh, 1997). In fact, top-down policies are not relevant for 21st century schools (Looney, 2009). Therefore, giving schools autonomy to translate TSLN is consistent with the principles of the policy.

However, the data suggest inconsistencies in the policy implementation between the macro and meso levels. In particular, while central policy on school change has been loosened, at the school level, procedures and processes have been tightened to ensure rigor and parity in the assessment tasks, and for accountability purposes. Hargreaves, Shirley, and Ng (2012, p. 80) refer to this misalignment as the "paradox of control" where there is "more autonomy, more control." They argue that while Singapore schools have been granted more autonomy from regulation and control, the public continues to hold the government responsible for ensuring high standards.

One example of this paradox of control is the teacher evaluation which serves to intensify the focus on performance and output (Liew, 2008). This instrument was intended to be both formative and summative. However, based on the teachers' comments in this study, schools tend to use the instrument summatively, and focus singularly on a quantitative measure—the MSG—to evaluate teachers despite the fact that teacher evaluation is comprised of a number of competencies (OECD, 2011). The effect was that teachers put a strong focus on performance targets (Liew, 2008).

To this end, it is important for policymakers to ensure that schools do not push teachers to enact teaching and assessment practices that do not reflect the TSLN intent of preparing students for the *test of life*. In such instances, it will be necessary for superintendents or policy makers to help the leadership teams of these schools understand the philosophies underlying the policy, to comprehend the reform goals, and to review ways to align their school-based processes and procedures with the macro vision.

Finally, there could be increased avenues for iterative policy making. As shown in Figure 7.1, the change process continues to move from the macro to the meso to the micro levels,

and based on the interview comments, the direction of change continues to emanate from the top. Such a change process contradicts the TSLN and TLLM goals of having schools be "crucibles" of change (C. T. Goh, 1997) and be places "bubbling" with ideas (Shanmugaratnam, 2005b). At the meso level, the teachers' comments also indicate the same top-down approach from the school leaders. Apart from Totoro and James who spoke of how they initiated new assessment types, the other teachers only mentioned taking direction from their heads of department. To this end, policy makers and school leaders could provide more open communication channels and revise policy based on consultation with and feedback from teachers. While the Our Singapore Conversation in 2012 was one such consultative effort,[34] the teachers' comments indicated that there continues to be top-down policy making. Perhaps, with reflexive and responsive policy making, the arrows in Figure 7.1 may morph from being one- to two-directional arrows.

Make explicit theoretical and philosophical underpinnings of the reform. This study suggests that in addition to communicating the intent, it will be useful for policymakers to be explicit about the undergirding principles and philosophies of the reform. This requires that the spirit of the reform be communicated clearly. While TSLN did not explicitly articulate an overarching philosophy, it is clear that the TLLM tenets draw on constructivist learning theories (Koh & Luke, 2009).

In order to realize the policy intent, decision makers have to be clear about the philosophical and theoretical underpinnings of the reform. On the one hand, the TLLM tenets provided teachers with the latitude to enact more of some strategies and less of others. On the other hand, the tendential nature of these tenets misleads teachers; there is no mandate that teachers have to use one or more of the strategies, nor are teachers evaluated for applying the

[34] This national conversation was mooted by the Prime Minister. See http://www.pmo.gov.sg/content/pmosite/mediacentre/speechesninterviews/primeminister/2012/August/prime_minister_leehsienloongsnationaldayrally2012speechinenglish.html#.UwxnnfmSxPI

tenets well. To this end, policy makers could be more explicit about the adoption of the tenets in classroom teaching, and this could be achieved by incorporating the tenets into curriculum materials. Another aspect that was missing was a compelling philosophy to frame the assessment-related tenets. This explains why the teachers discussed their assessment practices as a collection of strategies rather than point to an overarching assessment philosophy or principle.

The absence of an overarching assessment philosophy resulted in several teachers' classroom assessment practices reflecting the policy only to the letter. In addition, the practices converge toward closed questioning and focused on providing correct responses, approaches that resonate with behaviorist theories in which the teachers focus on teaching and assessing the next identified and planned concept or fact in a linear fashion (Torrance & Pryor, 2001). Students' response to such teaching and assessment approaches were to rely on their teachers for responses and solutions. They were not prompted or incited to energetically participate in the learning process or to figure out responses themselves.

In view of the TSLN intent to develop independent and engaged learners, it would be beneficial for policy makers to define the roles of teachers and students more explicitly within the overarching philosophy. In light of this, Singapore could take inspiration from Scotland's *Assessment is for Learning* (AifL) reform (discussed in Chapter 2) which envisions teachers and students working closely together in assessment activities. This joint teacher-student role in assessment is diagrammatically represented (see AifL Triangle in Learning and Teaching Scotland, 2006), and strongly signals a shift toward student-centered practices in which teachers allow students a degree of responsibility and ownership in the classroom (Cuban, 1993). In comparison to the Scottish experience, in Singapore, there is no role articulated for the student or the teacher, and this is one area that policy makers could consider.

Help teachers understand the theoretical underpinnings of the reform. Policy makers need to engage teachers and help them comprehend the philosophical and theoretical underpinnings of the reform. This suggestion draws on the research which posits that changes to teachers' conservatism in the classroom can and will only take place when teachers' cultural perceptions of teaching, learning, and assessment have been altered (e.g., Fullan, 2007; A. Hargreaves, et al., 2002; House, 1978, 1981). In this study, no more than three teachers' views on assessment and student learning embodied a cultural and philosophical alignment to the TSLN and TLLM intents. The other teachers, while agreeing with the TSLN goals, mostly held views of "assessment" and enacted assessment practices that resonated with a factory-model of education. In short, their views and practices only reflected the policy intent to the letter, but not to the spirit. Such enactment is "cultural morphing" which is a way teachers and stakeholders incorporate reforms within the existing cultural norms (C. L. Goh & Wong, 2013).

To bring about changed practices, policy makers will have to work with teacher-educators to help teachers comprehend and interpret their beliefs and values prior to and during a reform effort. Adopting constructivist approaches means allowing a larger student presence in the classroom, and this change in classroom dynamics may be uncomfortable for Singapore teachers for whom in teaching, "authority is hierarchy," and "classroom talk is teacher-dominated" (Hogan, 2014). Teaching practices in Singapore are embedded in the deep cultural traditions of the Asian region as a whole. For instance, research on Hong Kong's *Assessment for Learning* policy found that the policy-practice divergence in which teachers continued with teacher-led assessment practices instead of using student-centered approaches was due to the complexities of introducing a western assessment reform within an Asian society (Forrester & Wong, 2008). Therefore, to ensure more alignment to the spirit of the reform, teacher educators

and policy makers have to help teachers reconcile western-style approaches with traditional Asian views of teaching and learning (C. L. Goh & Wong, 2013).

When teachers are not acquainted with the overarching theoretical or philosophical underpinnings of classroom assessment in relation to the policy and to learning, their practices are less effective (Black & Wiliam, 2012b). These practices focus on procedures rather than on alignment with the spirit of the practice. In a similar vein, the absence of a theoretical underpinning in TLLM is one plausible reason why teachers' classroom assessment practices continue to be aligned with behaviorist rather than constructivist perspectives of learning. To this end, policy makers could work with researchers to make theory more accessible and available to classroom teachers, so that practices can become more aligned with the spirit of the principles underpinning policies and reforms.

Implications for research and teacher educators

Based on the meta-inferences discussed in this chapter, there are areas for research and of which teacher educators must be cognizant. The implications for research include the aspects discussed in this section, as well as those described under the "Limitations" section of this chapter.

Develop new ways to study student achievement. International benchmarking studies like TIMSS recognize the value of subjects like mathematics and science in preparing students to succeed during their time in formal education as well as to participate in daily life and in the workforce, and thus these studies provide countries with a means to measure progress in these two subjects (Mullis, et al., 2009). However, the drawback of student achievement from such data is that large-scale tests are limited in the extent to which they are able to assess a varied range of abilities (Atkin & Black, 1997). Given the arguments for the use of alternative

assessments that would engage students in the process of learning, it would be useful for researchers to work toward changing the nature of large-scale assessments to enable the assessment of a larger range of skills and cognitive domains. Furthermore, successful participation in society and in the workforce require students to be able to communicate their knowledge, arguments and ideas to a variety of audiences (Newmann & Associates, 1996) and in different ways via a range of media. To this end, researchers could explore test formats that require students to demonstrate their learning in more ways than just through multiple-choice questions and short answer constructed-response tasks which have been deemed as requiring only limited responses (Atkin & Black, 1997; Wiggins, 1992). One possible approach would be to re-surface performance assessments, such as that used for TIMSS 1995 which assessed a range of domains including procedural knowledge and problem-solving (Harmon et al., 1997). While such assessments enable students to be evaluated in life-like assessments, they are costly and complex to conduct and implement (Harmon, et al., 1997). However, in view of the changing nature of knowledge and the demands of society, it is necessary to establish measures that can assess student learning across a range of domains.

Understand educational change in Asian societies. First, in light of Singapore students' stellar performance in international benchmarking studies, it is evident that the curricular changes, such as the content cuts, made in response to TSLN have not put students at a disadvantage in terms of learning and meeting international standards. In view of the new assessment objectives introduced into the syllabuses, such as that analyzed for this study, research could explore ways to analyze teachers' assessment practices more deeply, particularly regarding how Asian countries adopt and adapt Western-based reforms to change their practices. There is a long historical Chinese culture of assessment in Singapore (Koh & Luke, 2009). The

high-stakes assessments that require students to reproduce learned material and that are widely used for individual selection originated from the imperial examinations in China (Madaus & O'Dwyer, 1999). Therefore, studying the attempts by Asian educational jurisdictions such as Singapore and Hong Kong to adopt and apply constructivist principles of teaching, learning, and assessment in the context of these historical traditions would provide important assessment lessons for the region, and for the world.

Examine teachers' thinking and reflections. Research on the examination of teachers' practices might also include more conversations with teachers to understand their beliefs and perspectives in response to new policies. Yet, many studies (e.g., those undertaken by the CORS and QSRLS teams, among others) have not typically involved teachers in reflecting on and discussing their professional perspectives on assessment practices. Through the interviews in this study, the reasons for teachers' assessment objectives and inquiry into how they enacted formative assessment to enhance learning provided deeper insight into their classroom assessment practices in relation to the policy intent. Further research could be conducted on a larger scale to analyze the decisions teachers make and the objectives they adopt when designing classroom assessments. The findings of such research would inform policy on how to best engage teachers and help them respond to the policy intent.

Provide sustained teacher education in assessment competencies. Professional development needs to be provided for teachers in order for them to enact new curricular and assessment practices well. While lectures and seminars can be provided to scale up and deepen teachers' assessment knowledge, a more critical concern pertains to the sustainability of the learning. Recent work reports that the quality of teacher assessments increased significantly when teachers participated in ongoing and continual professional learning in assessment literacy

(e.g., Koh, 2011b). For instance, over a two-year period, Koh (2011b) provided a series of professional learning sessions including guiding teachers to develop authentic assessment tasks and rubrics, and having them engage in peer critique and discussion. This kind of professional development model embodies the features of effective professional development (e.g., duration, coherence, content focus and collective participation) and is more powerful than short-term, once-off sessions (Koh, 2011b). Drawing on this model of sustained professional development, researchers and teacher educators must collaborate closely with policy makers to review current models of professional development, and to provide more localized and continuous professional learning in the area of classroom assessment practices. Such a partnership is possible in Singapore, which, unlike larger education systems, benefits from a close tripartite relationship among the Ministry of Education, the National Institute of Education (Singapore's teacher training institution) and schools (OECD, 2011). This partnership enables quick communication and implementation of policy changes (A. Hargreaves, et al., 2012; OECD, 2011).

Implications for practice in schools

Broadly, the macro data indicated that across all Singapore schools, students had teachers who implemented a variety of assessment types, and who frequently assessed higher-order skills. On the surface, these trends are aligned with the policy intent. However, when teachers' assessments were examined in greater depth, the micro data showed that the teachers did not always address higher-order skills extensively in the tasks they presented to their students. In fact, both the macro and micro data showed that teachers continued to require students to know factual knowledge and content. How then can classroom assessments be shifted further toward the "more" aspects of the TLLM tenets? While the TLLM tenets have already attempted to

change the classroom culture by repositioning the teacher as a facilitator and a guide, the roles of students and principals also need to be reviewed.

Augment the role of the student. Several teachers in this study lamented the fact that their students were quiet in class, usually waiting—unresponsively—for teachers to supply answers and solutions to problems and tasks. Yet, because the TSLN vision is for schools to nurture future citizens who are lifelong learners and independent thinkers, students need to participate actively in class and develop dispositions that make them curious and thirsty knowledge.

To realize TSLN's goals, there is a need to expand the role of the student in the teaching and learning process. Behaviorist assessment practices relegate students to being participants who simply wait for the teacher to dispense information, solutions, and strategies. Providing feedback using teacher-centered approaches can be an efficient way for teachers to overcome time constraints, but it also makes students overly dependent on the teacher for the solution or correct answer. Effective feedback, rather, must help students monitor, direct and regulate their actions in relation to the learning objectives (Hattie & Timperley, 2007).

To realize this, teachers need to "communicate a sense of partnership to students" by sharing learning goals with students (Kapambwe, 2010) and teaching them to gauge where they are in that learning process (Looney, 2011). This involves ongoing dialogue between teachers and students, and among students themselves as engaged and energetic contributors to their learning (A. Hargreaves, et al., 2002) as well as owners of that learning (Wiliam & Thompson, 2008).

One way to create more student ownership in learning is to involve students in self- and peer-assessment (Black & Wiliam, 1998a; Sadler, 1989). Self-assessment is beneficial because

it is ipsative in relating a student's previous achievement to the intended level (Weeden, Winter, & Broadfoot, 2002), and because it "places the student at the center of the assessment activity" (Lambert & Lines, 2000, p. 14). Self-assessment is a powerful way to make students aware of the progress they have made (Weeden, et al., 2002) and it helps them to develop meta-cognition skills (Black, et al., 2003b), so that they are reflective and cognizant of their thinking processes.

However, students who have been socialized to accept that there is one correct response told or given to them by the teacher, may resent or struggle with classroom activities and discourse that require them to be more participatory and to take responsibility for their learning (Black, et al., 2003b). To be aligned to the TSLN intent, teachers will need to help students manage and monitor their learning (Paige & Witty, 2008). Independent learners are those who want to approach teachers for assistance (Black, et al., 2003b), rather than ones who rely on teachers to provide solutions and answers.

These efforts to transform the role of the student require a culture of trust to be built in the classroom (Black & Wiliam, 2012b; Crooks, 1988; Stobart, 2008). The absence of trust and mutual respect in the classroom would result in students limiting the extent to which they disclose their thinking and ideas (Cowie, 2005). Unless students feel safe in in the classroom, they will remain quiet in class as they do in James' class (a *less aligned* teacher), and merely rely on their teachers to supply the necessary information.

Equip principals to be assessment leaders. The findings from this study also have implications for principals, given that they make decisions for schools' curricular programs and give approval to the types of support and resources teachers require in order to bring about student learning. The data provided evidence that some school level policies ran counter to the TSLN vision resulting in misalignment between the macro and meso levels. This was

particularly the case when schools made high-stakes use of the teacher evaluation tool or when schools instituted the standardization of assessment processes. The former caused teachers to be preoccupied with testing and retesting their students to generate higher marks and the latter resulted in teachers being less able to customize assessment tasks and formative assessments for students. There is therefore an urgent need for principals to be more mindful of the impact of school level policies on teachers' classroom assessment practices, especially if they detract from the policy intent. For instance, to enhance student learning, principals need to move beyond being administrative consumers of assessment data. Rather, they need to be equipped with assessment literacy skills so that they are able to lead their staff in interpreting, evaluating and analyzing assessment data together (Stiggins, 2001).

To bring about fundamental change in teachers' classroom practices, principals can provide assessment leadership in two areas. First, they should build the foundation for teachers' assessment literacy in their schools (Guskey, 2009). In view of the TSLN vision and TLLM tenets, principals must establish a balanced assessment system in their schools, one that combines formative and summative assessments purposefully and meaningfully (Jakicic, 2009). A balanced assessment system in schools would enable teachers to find out which students are learning, and how to help those who are struggling (Jakicic, 2009). Such a system would guide principals in deciding which curriculum and instruction approaches need changing (Jakicic, 2009). Adopting balanced assessment systems in schools would enable principals to steer teachers such as James and Marianne away from the emphasis on marks and achievement and to focus their efforts on helping students learn.

Another aspect of building an assessment foundation is for principals to establish a strong culture of learning in their schools (Erkens, 2009b), thereby leading teachers away from the

focus on achievement toward an emphasis on learning. A focus on learning means that principals work with their teachers to identify the intended learning goals and outcomes, and that they encourage their staff to adopt high expectations of students (Huff, 2009). Building a foundation for assessment literacy in the school would require principals to model assessment strategies (e.g., reflection, effective feedback), and to support teachers like Harry who want to try new assessment practices to challenge their students.

Second, as assessment leaders, principals need to work with their staff to collect, interpret, and report assessment data (Guskey, 2009). This involves principals being equipped with the skills to understand assessment results, and being able to create an open and safe environment for teachers to discuss and analyze student performance (Stiggins, 2001). With Singapore schools being urged to undertake more "qualitative" assessing such as using reflection logs and journals, teachers may receive a lot of assessment information which they need to decipher and to make follow-up curricular decisions. The role of principals would be to guide their staff in using this assessment data to inform teaching and learning. One approach is for principals to foster among the staff a culture of collaborative inquiry that would involve teachers in discussing the types of assessment information that they collect, the effectiveness of particular curricular programs, and the ways they could proceed in identifying students' strengths and weakness (Vagle, 2009). In addition, with assessment data that has been collected, teachers could then examine the challenges they face in assessing and supporting student learning, and could identify the causes, and thereby select appropriate strategies to address the issues (Vagle, 2009). Such conversations would enable the staff to understand that they can assess and support learning in different ways (Vagle, 2009). Engaging teachers in inquiry and problem-solving is aligned to the TSLN intent of encouraging innovation and ideas on the ground. When enacted in schools, teachers working

together to examine assessment information is a means of changing the one-directional meso-micro relationship into a two-directional flow.

Promote culture of collaboration. This study found that the teachers whose practices reflected the policy intent collaborated with and learned more from colleagues. These professional practices provided them with feedback and ideas to construct and improve their prompts and tasks. Therefore, one implication arising from this study is to encourage more teachers to participate in professional communities to improve their assessment practices rather than to struggle in isolation. Professional collaboration can be formal or informal. In terms of formal arrangements, there was evidence of teachers working together to align the lower and upper secondary syllabuses (e.g., Jiajia's school). In terms of informal collaboration, Harry was a participant in a community of teachers, working to improve his assessment practices through inter-departmental work. Just as teachers collaborate to design lessons and discuss instructional activities, teachers can also participate in learning communities to examine assessment practices.

Transformed cultures of professional collaboration would see teachers participating and engaging in deep conversations about assessment practices with colleagues, an activity which would provide them with valuable ideas and strategies to improve their assessment practices (Stiggins & Bridgeford, 1985). Through collaboration in the assessment design process, teachers could assist each other in constructing tasks, asking challenging questions, and lending a critical eye to review tests and assessment tasks. In these professional communities, teachers would be able to develop shared understanding of standards, to improve mastery of disciplinary knowledge and skills, to determine the necessary skills and knowledge that students should be learning, and to identify the types of evidence to ascertain what students have learned (Paige & Witty, 2008). As a team of professionals working together, teachers could engage in and promote "collective

autonomy" which conceives of teachers initiating change among colleagues (A. Hargreaves & Shirley, 2012, p. 197), without having to wait for directives from middle managers or school leaders. This aptly reflects the essence of TSLN because *thinking schools* were to be places of ideas and creativity (C. T. Goh, 1997).

One aspect that was missing from the teachers' discussion of professional collaboration, however, was working in concert to interpret assessment data. The teachers in the qualitative component of this study spoke at length about how they constructed test items, prompts, and other administrative strategies, yet the diagnosis of students' learning needs and analysis of students' performance appeared to be borne by each teacher individually. Research has reported benefits to teaching and learning when teachers work in teams to interpret and analyze assessment data (Datnow, Park, & Wohlstetter, 2007), and this is one practice that could be perpetrated in Singapore schools.

Provide more curriculum time. The findings from this study, as with the existing research (see also Gleeson, 2011; Yu & Frempong, 2012) point to the need for teachers to have more time to conduct their classroom assessments. A heavy marking load and limited curriculum time resulted in many of the teachers providing feedback to students in a factory-model manner in which formative assessment is standardized and mass produced in the form of short comments or model answers. Teachers had little time to offer individualized feedback on each student's strengths and weaknesses, and to provide strategies and approaches for improvement. Currently, customized feedback is only provided for upper secondary students or for higher-ability classes. However, since TSLN is situated within the ability-based, aspiration-driven paradigm, there is a need to customize formative assessment more to the needs of the individual student, regardless of grade level. One way of supporting teachers to fulfill these

purposes of better formative assessment would be to lessen or restructure their professional load to provide more time for personalized feedback to lower secondary students.

Summary

Drawing from the macro and micro data findings that there has been incremental change in teachers' assessment practices, the implications for policy, for research and teacher education, as well as for direct practice in schools are presented with the intent to shift practices from the incremental to the fundamental, and to suggest ways for more iterative and reflexive interactions across the macro, meso and micro levels. To bring about classroom assessment practices that are aligned to the TSLN vision and the TLLM tenets requires greater macro policy alignment, more support from research and teacher education, increased teacher collaboration and professional learning, a higher level of student participation in the classroom, and greater involvement by the principal as an assessment leader.

Preparing students for the *test of life*

Students who are ready to face the realities of the world and the 21st century when they leave school need to be comfortable with "ambiguities," to quote Totoro. This teacher's compelling and insightful observation speaks volumes about how and what teaching, learning and assessment must look like in the classroom if teachers are to prepare students for the *test of life*. Since life in the 21st century will be complex, volatile, uncertain, and ever-changing, teachers need to prepare their students to deal with unfamiliar situations and contexts, so that they will not be fazed when confronted with a myriad of situations and scenarios. Aligned to these objectives, classroom assessments that will equip students with the types of skills needed for life outside of school must require them to critique and analyze information, as well as to

apply learned knowledge to real world contexts (Biggs, 1995; Newmann & Associates, 1996; Wiggins, 1990).

Enacting assessment practices that make students ready for the *test of life* involves a complex interaction of personal, cultural, professional, administrative, and professional factors. This requires even closer links, coordination, and planning among policy makers, curriculum leaders, teacher educators, and researchers. In particular, planners and researchers have to work in concert with teachers to understand their beliefs about learning and education (Cuban, 1993). They also have to help teachers reconcile their perspectives and experiences with the intent of the policy reforms. In the interest of supporting each student in Singapore's ability-driven paradigm, the combined efforts of all parties will contribute to realizing the vision of preparing students to participate purposefully and meaningfully in the *test of life* when they graduate from school.

REFERENCES

Andrade, H. L. (2010). Summing up and moving forward: Key challenges and future directions for research and development in formative assessment. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 344-351). New York: Routledge.

Archbald, D. A., & Newmann, F. M. (1988). *Beyond standardized testing: Assessing authentic academic achievement in the secondary school*. Reston, VA: National Association of Secondary School Principals.

Aschbacher, P. R. (1999). Developing indicators of classroom practice to monitor and support school reform (CSE Technical Report 513). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles.

Aschbacher, P. R., & Alonzo, A. (2006). Examining the utility of elementary science notebooks for formative assessment purposes. *Educational Assessment, 11*(3-4), 179-203.

Assessment Reform Group. (1999). Assessment for learning: Beyond the black box (pp. 1-12). Cambridge, UK: University of Cambridge School of Education.

Assessment Reform Group. (2002). Assessment for Learning: Beyond the black box. Retrieved November 10, 2008, from http://www.assessment-reform-group.org/CIE3.PDF

ATC21s.org. (2011, 13 November 2011). Transforming Education: Assessing and Teaching 21st Century Skills Retrieved 13 November, 2011, from http://atc21s.org/wp-content/uploads/2011/04/Cisco-Intel-Microsoft-Assessment-Call-to-Action.pdf

Atkin, J. M., & Black, P. (1997). Policy perils of international comparisons: The TIMSS case. *Phi Delta Kappan, 79*(1), 22-28.

Avery, P. G. (1999). Authentic assessment and instruction. *Social Education, 63*(6), 368-373.

Avery, P. G., Freeman, C., & Carmichael-Tanaka, D. (2002). Developing authentic instruction in the social studies. *Journal of Research in Education, 12*(1), 50-56.

Avery, P. G., Jouneski, N. P., & Odendahl, T. (2001). Authentic pedagogy seminars: Renewing our commitment to teaching and learning. *The Social Studies, 92*(3), 97-101.

Ayala, C. C., Shavelson, R. J., Ruiz-Primo, M. A., Brandon, P. R., Yin, Y., Furtak, E. M., . . . Tomita, M. K. (2008). From formal embedded assessments to reflective lessons: The development of formative assessment studies. *Applied Measurement in Education, 21*(4), 315-334.

Beaton, A. E., Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., Smith, T. A., & Kelly, D. L. (1996). *Science achievement in the middle school years: IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Bell, B., & Cowie, B. (2001). The characteristics of formative assessment in science education. *Science Education, 85*(5), 536-553.

Ben-Chiam, D., Keret, Y., & Ilany, B.-S. (2007). Designing and implementiing authentic investigative proportional reasoning tasks: The impact on pre-service mathematics teachers' content and pedagogical knowledge and attitudes. *Journal of Mathematics Teacher Education, 10*(4-6), 333-340.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18*(1), 5-25.

Berends, M. (2006). Survey methods in educational research. In J. L. Green, G. Camilli & P. B. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 623-640). Mahwah, NJ: Lawrence Erlbaum Associates.

Berry, R. (2011). Assessment trends in Hong Kong: Seeking to establish formative assessment in an examination culture. *Assessment in Education: Principles, Policy & Practice, 18*(2), 199-211.

Biggs, J. B. (1992). A qualitative approach to grading students. *HERDSA News, 14*(3), 1-8.

Biggs, J. B. (1995). Assessing for learning: Some dimensions underlying new approache to educational assessment. *The Alberta Journal of Educational Research, XLI*(1), 1-17.

Biggs, J. B. (1996a). Assessing learning quality: Reconciling instutional, staff and educational demands. *Assessment & Evaluation in Higher Education, 21*(1), 5-16.

Biggs, J. B. (1996b). Western misperceptions of the Confucian-Heritage learning culture. In D. A. Watkins & J. B. Biggs (Eds.), *The Chinese learner: Cultural, psychological and contextual influences* (pp. 46-67). Hong Kong: Comparative Education Research Centre and the Australian Council for Educational Research.

Bishop, K. (2000). The research processes of gifted students: A case study. *Gifted Child Quarterly, 44*(1), 54-64.

Black, P. (1998). *Testing: friend or foe? The theory and practice of assessment and testing*. London ; Washington, D.C.: Falmer Press.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003a). *Assessment for learning : Putting it into practice*. Maidenhead: Open University Press.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003b). *Assessment for learning: Putting it into practice*. Berkshire, England: Open University Press.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working Inside the Black Box: Assessment for Learning in the Classroom. (cover story). *Phi Delta Kappan, 86*(1), 9-21.

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7-74. doi: 10.1080/0969595980050102

Black, P., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, October*, 139-148.

Black, P., & Wiliam, D. (1998c). *Inside the black box: Raising standards through classroom assessment*. London: King's College London.

Black, P., & Wiliam, D. (2003). "In praise of educational research": Formative assessment. *British Educational Research Journal, 29*(5), 623-637.

Black, P., & Wiliam, D. (2005). Lessons from around the world: How policies, politics and cultures constrain and afford assessment practices. *The Curriculum Journal, 16*(2), 249-261.

Black, P., & Wiliam, D. (2006). Developing a theory of formative assessment. In J. Gardner (Ed.), *Assessment and learning* (pp. 81-100). London: Sage.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation, and Accountability, 21*(1), 5-31. doi: 10.1007/s11092-008-9068-5

Black, P., & Wiliam, D. (2012a). Assessment for learning in the classroom. In J. Gardner (Ed.), *Assessment and learning* (pp. 11-32). London, UK: Sage.

Black, P., & Wiliam, D. (2012b). Developing a theory of formative assessment. In J. Gardner (Ed.), *Assessment and learning* (pp. 206-229). London, UK: Sage.

Bloom, B., Hastings, J. T., & Madaus, G. F. (Eds.). (1971). *Handbook on the formative and summative evaluation of student learning*. New York: McGraw-Hill.

Bol, L., Stephenson, P. L., O'Connell, A. A., & Nunnery, J. A. (1998). Influence of experience, grade level, and subject area of teachers' assessment practices. *Journal of Educational Research, 91*(6), 323-330.

Bol, L., & Strage, A. (1996). The contradiction between teachers' instructional goals and their assessment practices in high school biology courses. *Science Education, 80*(2), 145-163.

Boyle, W. F., & Charles, M. (2010). Leading through Assessment for Learning? *School Leadership and Management, 30*(3), 285-300.

Braden, J. P., Schroeder, J. L., & Buckley, J. A. (2001). Secondary school reform, inclusion, and authentic assessment. Madison, WI: Wisconsin Center for Education Research.

Bransford, J., Brown, A. L., & Cocking, R. (2000). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.

Bredo, E. (2006). Philosophies of Educational Research. In J. L. Green, G. Camilli & P. B. Elmore (Eds.), *Handbook of complementary methods in education research*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Brenner, M. E. (2006). Interviewing in educational research. In J. L. Green, G. Camilli & P. B. Elmore (Eds.), *Handbook of complementary methods in education research*. Mahwah, NJ: Lawrence Erlbaum Associates.

Broadfoot, P., & Black, P. (2004). Redefining assessment? The first ten years of assessment in education. *Assessment in Education: Principles, Policy & Practice, 11*(1), 7-26. doi: 10.1080/0969594042000208976

Brookhart, S. M. (2001). Successful students' formative and summative uses of assessment information. *Assessment in Education: Principles, Policy & Practice, 8*(2), 153-169.

Brookhart, S. M. (2004). Classroom assessment: Tensions and intersections in theory and practice. *Teachers College Record, 106*(3), 429-458.

Brookhart, S. M. (2007). Expanding views about formative classroom assessment: A review of the literature. In J. H. McMillan (Ed.), *Formative classroom assessment: Theory into practice* (pp. 43-62). New York: Teachers College Press.

Brooks, J. G., & Brooks, M. G. (1993). *In search of understanding: The case for constructivist classrooms*. Alexandria, VA: Association for Supervision and Curriculum Development.

Brown, G. T. L. (2004). Teachers' conceptions of assessment: Implications for policy and professional development. *Assessment in Education: Principles, Policy & Practice, 11*(3), 301-318.

Brown, G. T. L., Kennedy, K. J., Fok, P. K., Chan, J. K. S., & Yu, W. M. (2009). Assessment for student improvement: Understanding Hong Kong teachers' conceptions and practices of assessment. *Assessment in Education: Principles, Policy & Practice, 16*(3), 347-363.

Bryk, A. S., Nagaoka, J. K., & Newmann, F. M. (2000). Chicago classroom demands for authentic intellectual work: Trends from 1997-1999 (pp. 1-14). Chicago, IL: Consortium on Chicago School Research.

Callison, D., & Lamb, A. (2004). Authentic learning. *School Library Media Activities Monthly, XXI*(4), 34-39.

Camp, R. (1992). Assessment in the context of schools and school change. In H. H. Marshall (Ed.), *Redefining student learning* (pp. 241-263). Norwood, NJ: Ablex Publishing Corporation.

Carless, D. (2005). Prospects for the implementation of assessment for learning. *Assessment in Education: Principles, Policy & Practice, 12*(1), 39-54.
Carless, D. (2007). Conceptualizing Pre-emptive Formative Assessment. *Assessment in Education: Principles, Policy & Practice, 14*(2), 171-184.
Cassidy, K. E. (2009). *Using authentic intellectual assessment to determine level of instructional quality of teacher practice of new elementary school teacenrs based on teacher preparation route*. Unpublished dissertation thesis, Graduate School of Education and Human Development, George Washington University. Washington, D.C.
Channelnewsasia. (2012). MOE introduces new component in lower secondary humanities subjects Retrieved 25 February, 2012, from http://www.channelnewsasia.com/stories/singaporelocalnews/view/1181941/1/.html
Charmaz, K. (2000). Grounded theory: Objectivist and constructivist methods. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 509-535). Thousand Oaks, CA: Sage.
Chin, C., & Teou, L.-Y. (2009). Using concept cartoons in formative assessment: Scaffolding studnets' argumentation. *International Journal of Science Education, 31*(10), 1307-1332. doi: 10.1080/09500690801953179
Christenson, S. L. (1991). Authentic assessment: Straw man or prescription for progress. *School Psychology Quarterly, 6*(4), 294-299.
Cizek, G. J. (1997). Learning, achievement, and assessment: Constructs at a crossroads. In G. D. Phye (Ed.), *Handbook of classroom assessment: Learning, achievement and adjustment* (pp. 1-32). San Diego, CA: Academic Press.
Cizek, G. J. (2010). An introduction to formative assessment: History, characteristics, and challenges. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 3-17). New York: Routledge.
Clare, L. (2000). Using teachers' assignments as an indicator of classroom practice (CSE Technical Report 532). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles.
Clare, L., & Aschbacher, P. R. (2001). Exploring the technical quality of using assignments and student work as indicators of classroom practice. *Educational Assessment, 7*(1), 39-59.
Clare, L., Valdes, R., Pascal, J., & Steinberg, J. R. (2001). Teachers' assignments as indicators of instructional quality in elementary schools (CSE Technical Report 545). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles.
Clark, I. (2011). Formative assessment: Policy, perspectives and practice. *Florida Journal of Educational Administration & Policy, 4*(2), 158-180.
Cochran-Smith, M., Shakman, K., Jong, C., Terrell, D. G., Barnatt, J., & McQuillan, P. J. (2009). Good and just teaching: The case for social justice in teacher education. *American Journal of Education, 115*(3), 347-377.
Coffey, J. E., Sato, M., & Thiebault, M. (2005). Classroom assessment up close and personal. *Teacher Development, 9*(2), 169-184.
Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education*. London: Routledge.
Colby-Kelly, C., & Turner, C. E. (2007). AFL research in the L2 classroom and evidence of usefulness: Taking formative assessment to the next level. *Canadian Modern Language Review, 64*(1), 9-37.

Cole, N. S. (1990). Conceptuations of educational achievement. *Educational Researcher, 19*(3), 2-7.
Collins Cobuild Advanced Learner's English Dictionary. (Ed.) (2005). Glasgow, U.K.: HarperCollins.
Common Core State Standards Initiative. (2010). Reactions to the March 2010 Draft Common Core State Standards: Highlights and themes from public feedback. Retrieved October 28, 2010, from http://www.corestandards.org/assets/k-12-feedback-summary.pdf
Cooper, B., & Cowie, B. (2010). Collaborative research for assessment for learning. *Teaching and Teacher Education, 26*(4), 979-986.
Cowie, B. (2005). Pupil commentary on assessment for learning. *The Curriculum Journal, 16*(2), 137-151.
Cowie, B., & Bell, B. (1999). A model of formative assessment in science education. *Assessment in Education: Principles, Policy & Practice, 6*(1), 101-116.
Creswell, J. W. (2008). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (3rd ed.). Upper Saddle River, NJ: Pearson Education.
Creswell, J. W. (2010). Mapping the developing landscape of mixed methods research. In A. Tashakkori & C. Teddlie (Eds.), *Sage handbook of mixed methods research in social behavioral research* (2nd ed., pp. 45-68). Thousand Oaks, CA: Sage.
Creswell, J. W., & Plano Clark, V. L. (2007). *Design and conducting mixed methods research*. Thousand Oaks, CA: Sage.
Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58*(4), 438-481.
Crossouard, B. (2009). A sociocultural reflection on formative assessment and collaborative challenges in the state of Jersey. *Research Papers in Education, 24*(1), 77-93. doi: 10.1080/13669870801945909
Crossouard, B. (2011). Using formative assessment to support complex learning in conditions of social adversity. *Assessment in Education: Principles, Policy & Practice, 18*(1), 59-72.
Cuban, L. (1984). *How teachers taught: Constancy and change in American classrooms, 1890-1980*. New York: Longman.
Cuban, L. (1993). *How teachers taught: Constancy and change in American classrooms, 1890-1990*. New York: Longman.
Cumming, J. J., & Maxwell, G. S. (1999). Contextualizing authentic assessment. *Assessment in Education: Principles, Policy & Practice, 6*(2), 177-194.
Curriculum Planning and Development Division. (1992). *Science syllabus. Lower seconary(Special/ Express/ Normal course)*. Singapore: Curriculum Planning Division, Ministry of Education.
Curriculum Planning and Development Division. (2000). *Science syllabus: Lower secondary (Special/Express, Normal Academic)*. Singapore: Curriculum Planning & Development Division, Ministry of Education.
Curriculum Planning and Development Division. (2005). *Geography syllabus: Lower secondary. [Express/Special, Normal (Academic)]*. Singapore: Curriculum Planning & Development Division, Ministry of Education.
Curriculum Planning and Development Division. (2007). *2008 Syllabus Science-Lower Secondary [Express/ Normal (Academic)]*. Singapore: Curriculum Planning & Development Division, Ministry of Education.

D'Agostino, J. V. (1996). Authentic instruction and academic achievement in compenstaory education classrooms. *Studies in Educational Evaluation, 22*(2), 139-155.

Dahlin, B., Watkins, D. A., & Ekholm, M. (2001). The role of assessment in student learning: The views of Hong Kong and Swedish lecturers. In D. A. Watkins & J. B. Biggs (Eds.), *The Chinese learner: Psychological and pedagogical perspectives.* (pp. 47-74). Hong Kong: Comparative Education Research Centre, the University of Hong Kong.

Darling-Hammond, L. (2010). New policies for 21st century demands. In J. Bellenca & R. Brandt (Eds.), *21st century skills: Rethinking how students learn* (pp. 32-49). Bloomington, IN: Solution Tree.

Darling-Hammond, L., Ancess, J., & Falk, B. (1995). *Authentic assessment in action: Studies of schools and students at work*. New York: Teachers College Press.

Datnow, A., Park, V., & Wohlstetter, P. (2007). Achieving with data: How high-performing school systems use data to improve instruction for elementary students Retrieved 2 September, 2008, from http://www.usc.edu/dept/education/cegov/focus/education-reform/publications/books-chapters/Achieving%20with%20Data-How%20High%20Performing%20Schools%20Use%20Data%5B1%5D.pdf

Davies, P., Durbin, C., Clarke, J., & Dale, J. (2004). Developing students' conceptions of quality in geography. *The Curriculum Journal, 15*(1), 19-34.

Dekker, T., & Feijs, E. (2005). Scaling up strategies for change: Change in formative assessment practices. *Assessment in Education: Principles, Policy & Practice, 12*(3), 237-254.

Dennis, J., & O'Hair, M. J. (2010). Overcoming obstacles in using authentic instruction: A comparative case study of high school math and science teachers. *American Secondary Education, 38*(2), 4-22.

Desimone, L. M. (2009). Complementary methods for policy research. In G. Sykes, B. Schneider & D. N. Plank (Eds.), *Handbook of education policy research* (pp. 163-175). New York: Routledge & AERA.

Dibu-Ojerinde, O. O. (2005). Formative assessment for fearning. *International Journal of Learning, 12*(8), 355-360.

Dixon, H., & Haigh, M. (2009). Changing mathematics teachers' conceptions of assessment and feedback. *Teacher Development, 13*(2), 173-186.

Doering, A., & Veletsianos, G. (2008). An investigation of the use of real-time, authentic geospatial data in the K-12 classroom. *Journal of Geography, 106*(6), 217-225.

Dr Goh Keng Swee and the Education Study Team Ministry of Education. (1978). *Report on the Ministry of Education*. Singapore: Ministry of Education.

Eisner, E. W. (1991). Taking a second look: Educational connoisseurship revisited. In M. W. McLaughlin & D. C. Phillips (Eds.), *Evaluation and education: At quarter century* (pp. 168-187). Chicago: University of Chicago Press.

Elliott, S. N. (1991). Authentic assessment: An introduction to a neobehavioral approach to classroom assessment. *School Psychology Quarterly, 6*(4), 273-278.

Elmore, R. F. (2004). *School reform the inside out: Policy, practice and performance*. Cambridge, MA: Harvard Education Press.

Erkens, C. (2009a). Developing our assessment literacy. In T. R. Guskey (Ed.), *Teacher as assesment leader* (pp. 11-30). Bloomington, IN: Solution Tree Press.

Erkens, C. (2009b). Paving the way for an assessment-rich culture. In T. R. Guskey (Ed.), *The principal as assessment leader* (pp. 9-28). Bloomington, IN: Solution Tree Press.

Ertmer, P. A., & Newby, T. J. (1993). Behaviorism, cognitivism, constructivism: Comparing critical features from an instructional design perspective. *Performance Improvement Quarterly, 64*(4), 50-72.

European Parliament and Council of the European Union. (2006). Key competencies for lifelong learning, L394/10 C.F.R. Retrieved 30 September, 2011, from http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2006:394:0010:0018:en:PDF

Evans, J., & Benefield, P. (2001). Systematic reviews of educational research: Does the medical model fit? *British Educational Research Journal, 27*(5), 527-541. doi: 10.1080/01411920120095717

Feldman, A., & Capobianco, B. M. (2008). Teacher learning of technology enhanced formative assessment. *Journal of Science Education and Technology, 17*(1), 82-99.

Fischer, C. F., & King, R. M. (1996). *Authentic assessment: A guide to implementation*. Thousand Oaks, CA: Corwin Press.

Forrester, V., & Wong, M. (2008). Curriculum reform in the Hong Kong primary classroom: What gives? *Music Education Research, 10*(2), 271-284. doi: 10.1080/14613800802079122

Fox-Turnbull, W. (2006). The influences of teacher knowledge and authentic formative assessment on student learning in technology education. *International Journal of Technology & Design Education, 16*(1), 53-77.

Foy, P., & Olson, J. F. (Eds.). (2009). *TIMSS 2007 User guide for the international database*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Fraenkel, J. R., & Wallen, N. E. (2009). *How to design and evaluate research in education* (7th ed.). New York: McGraw-Hill.

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research, 74*(1), 59-109.

Freire, P. (2000). *Pedagogy of the oppressed*. New York: Continuum. (Original work published 1970).

Frey, B. B., & Schmitt, V. L. (2007). Coming to Terms with Classroom Assessment. *Journal of Advanced Academics, 18*(3), 402-423.

Fullan, M. (2007). *The new meaning of educational change* (4th ed.). New York: Teachers College Press.

Furtak, E. M., Ruiz-Primo, M. A., Shemwell, J. T., Ayala, C. C., Brandon, P. R., Shavelson, R. J., & Yin, Y. (2008). On the fidelity of implementing embedded formative assessments and its relation to student learning. *Applied Measurement in Education, 21*(4), 360-389.

Gardner, J. (Ed.). (2006). *Assessment and learning*. London: Sage.

Gareis, C. R. (2007). Reclaiming an important teacher competency: The lost art of formative assessment. *Journal of Personnel Evaluation in Education, 20*(1-2), 17-20.

Gattullo, F. (2000). Formative assessment in primary (elementary) ELT classes: An Italian case study. *Language Testing, 17*(2), 278-288.

Gioka, O. (2006). Assessment for learning in physics investigations: Assessment criteria, questions and feedback in marking. *Physics Education, 41*(4), 342-346.

Gioka, O. (2009). Teacher or examiner? The tension between formative and summative assessment in the case of science coursework. *Research in Science Education, 39*(4), 411-428.

Gipps, C. (1999). Socio-cultural aspects of assessment. *Review of Research in Education, 24*, 355-392.
Gipps, C. (2002). Sociocultual perspectives on assessment. In G. Wells & G. Claxton (Eds.), *Learning for life in the 21st century* (pp. 73-83). Oxford, UK: Blackwell.
Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Chicago: Aldine Publishing.
Gleeson, A. M. (2011). *Preparing teachers and students for democracy: Teachers and student leanring and authentic intellectual work*. Unpublished doctoral thesis, Lynch School of Education, Boston College. Boston, MA.
Goh, C. L., & Wong, T. (2013, October 5). Bridging the education gap, *The Straits Times,* pp. D2-D3.
Goh, C. T. (1997). Shaping our future: Thinking Schools, Learning Nation. Speech by Prime Minister Goh Chok Tong at the opening of the 7th international conference on thinking on Monday, 2 June 1997 at the Suntec City Convention Centre Ballroom Retrieved 25 September, 2011, from http://www.moe.gov.sg/media/speeches/1997/020697.htm
Goh, C. T. (1999). Prime Minister's National Day Rally Speech, 1999: First world economy, world-class home - Extract E. Education. Retrieved 1 February, 2013, from http://www.moe.gov.sg/media/speeches/1999/sp270899.htm
Good, R. (2011). Formative use of assessment information: It's a process, so let's say what we mean. *Practical Assessment, Research and Evaluation, 16*(3), 1-6. Retrieved from http://pareonline.net/pdf/v16n3.pdf
Good, T. L., & Brophy, J. E. (2008). *Looking into classrooms* (10th ed.). Boston, MA: Pearson Education.
Gopinathan, S. (1999). Preparing for the next rung: Economic restructuring and educational reform in Singapore. *Journal of Education and Work, 12*(3), 295-308.
Gopinathan, S. (2007). Globalisation, the Singapore developmental state and education policy: A thesis revisited. *Globalisation, Societies and Education, 5*(1), 53-70. doi: 10.1080/147720601133405
Grant, S. G., Gradwell, J. M., & Cimbricz, S. K. (2004). A question of authenticity: The document-based question as an assessment of studnets' knowledge of history. *Journal of Curriculum and Supervision, 19*(4), 309-337.
Graue, M. E. (1993). Integrating theory and practice through instructional assessment. *Educational Assessment, 1*(4), 293-309.
Greene, J. C. (2001). Mixing social inquiry methodologies. In V. Richardson (Ed.), *Handbook of research on teaching* (pp. 251-258). Thousand Oaks, CA: Sage.
Greene, J. C., Benjamin, L., & Goodyear, L. (2001). The merits of mixing methods in evaluation. *Evaluation, 7*(1), 25-44.
Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Towards a conceptual framework for mixed-methd evaluation designs. *Educational Evaluation and Policy Analysis, 11*(3), 255-274.
Gresham, F. M. (1991). Alternative psychometrics for authentic assessment. *School Psychology Quarterly, 6*(4), 305-309.
Grisham-Brown, J., Hallam, R. A., & Pretti-Frontczak, K. (2008). Preparing Head Start personnel to use a curriculum-based assessment: An innovative practice in the "age of accountability". *Journal of Early Intervention, 30*(4), 271-281.

Gulikers, J., Bastiens, T. J., & Krischner, P. A. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research and Development, 52*(3), 67-86.
Guskey, T. R. (2003). How classroom assessments improve learning. *Educational Leadership, 60*(5), 6-11.
Guskey, T. R. (2009). Introduction. In T. R. Guskey (Ed.), *The principal as assessment leader* (pp. 1-6). Bloomington, IN: Solution Tree Press.
Hairon, S. (2008). Teacher professional development in the TSLN era: Current challenges and future directions. In J. Tan & P. T. Ng (Eds.), *Thinking schools, learning nation: Contemporary issues and challeges*. Singapore: Pearson Education South Asia.
Hamilton, L. S., & Berends, M. (2006). *Instructional practices related to standards and assessments*. Paper presented at the American Educational Research Association Annual Meeting, San Francisco, CA. Retrieved from http://www.rand.org/pubs/working_papers/2006/RAND_WR374.pdf
Hammersley, M. (2001). On 'systematic' reviews of research literatures: A 'narrative' response to Evans and Benefield. *British Educational Research Journal, 27*(5), 543-554. doi: 10.1080/3054980020001882
Hanley-Maxwell, C., Phelps, L. A., Braden, J. P., & Warren, V. (1999). Schools of authentic and inclusive learning. Madison, WI: Wisconsin Center for Education Research.
Hargreaves, A. (1994). *Changing teachers, changing times*. New York: Teachers College Press.
Hargreaves, A. (2003). *Teaching in the knowledge society: Education in the age of insecurity*. New York: Teachers College Press.
Hargreaves, A., Earl, L., & Schmidt, M. (2002). Perspectives on alternative assessment reform. *American Educational Research Journal, 39*(1), 69-95.
Hargreaves, A., & Fullan, M. (2012). *Professional capital: Transforming teaching in every school*. New York: Teachers College Press.
Hargreaves, A., & Goodson, I. (2006). Educational change over time? The sustainability and nonsustainability of three decades of secondary school change and continuity. *Educational Administration Quarterly, 42*(1), 3-41.
Hargreaves, A., & Shirley, D. (2012). *The global fourth way: The quest for educational excellence*. Thousand Oaks, CA: Corwin.
Hargreaves, A., Shirley, D., & Ng, P. T. (2012). Singapore: Innovation, communication, and paradox. In A. Hargreaves & D. Shirley (Eds.), *The global fourth way: The question for educational excellence* (pp. 71-91). Thousand Oaks, CA: Sage.
Hargreaves, E. (2005). Assessment for learning? Thinking outside the (black) box. *Cambridge Journal of Education, 35*(2), 213-224.
Harlen, W. (2009). Assessment for learning: Research that is convincing (Part 1). *Education in science: The bulletin of the Association for Science Education, 231*(February), 30-31.
Harlen, W., & James, M. (1997). Assessment and learning: Differences and relationships between formative and summative assessment. *Assessment in Education: Principles, Policy & Practice, 4*(3), 365-379.
Harmon, M., Smith, T. A., Martin, M. O., Kelly, D. L., Beaton, A. E., Mullis, I. V. S., . . . Orpwood, G. (1997). *Performance assessment in IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, MA: TIMSS International Study Center, Boston College.
Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.

Hattie, J., & Timperley, H. S. (2007). The power of feedback. *Review of Educational Research, 77*(1), 88-112. doi: 10.3102/003465430298487

Hayes, D., Mills, M., Christie, P., & Lingard, B. (2006). *Teachers and schooling making a difference: Productive pedagogies, assessment and performance*. Crows Nest, NSW, Australia: Allen & Unwin.

Hayward, L. (2007). Curriculum, pedagogies and assessment in Scotland: The quest for social justice. 'Ah kent yir faither'. *Assessment in Education: Principles, Policy & Practice, 14*(2), 215-268.

Hayward, L., & Hedge, N. (2005). Travelling towards change in assessment: policy, practice and research in education. *Assessment in Education: Principles, Policy & Practice, 12*(1), 55-75.

Hayward, L., & Spencer, E. (2010). The complexities of change: Formative assesment in Scotland. *The Curriculum Journal, 21*(2), 161-177.

Heng, S. K. (2011). Our children. Our Purpose. Our future. Speech presented by Mr Heng Swee Keat, Minister for Education, at the Ministry of Education (MOE) Work Plan Seminar 2011 at the Ngee ann Polytechnic Convention Centre on Thursday, 22 September 2011. Retrieved 23 September, 2011, from http://www.moe.gov.sg/media/speeches/2011/09/22/work-plan-seminar-2011.php

Heng, S. K. (2012). Keynote address by Mr Heng Swee Keat, Minister for Education, at the Ministry of Education Work Plan Seminar, on Wednesday, 12 September 2012 at 9.20 a.m. at Ngee Ann Polytechnic Convention Centre. Retrieved 1 November, 2012, from http://www.moe.gov.sg/media/speeches/2012/09/12/keynote-address-by-mr-heng-swee-keat-at-wps-2012.php

Herman, J. L., Aschbacher, P. R., & Winters, L. (1992). *A practical guide to alternative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

Hesse-Biber, S. N. (2010). *Mixed methods research: Merging theory with practice*. New York: Guildford Press.

Hirsch, E. D. (1987). *Cultural literacy: What every American needs to know*. New York: Houghton-Mifflin Co.

Hodgen, J., & Marshall, B. (2005). Assessment for learning in English and mathematics: A comparison. *The Curriculum Journal, 16*(2), 153-176.

Hogan, D. (2014). Why is Singapore's school system so successful, and is it a model for the West? Retrieved 12 February, 2014, from http://theconversation.com/why-is-singapores-school-system-so-successful-and-is-it-a-model-for-the-west-22917

House, E. R. (1978). Technology versus craft: A ten year perspective on innovation. In P. H. Taylor (Ed.), *New directions in curriculum studies* (pp. 137-151). London: Falmer Press.

House, E. R. (1981). Three perspectives on innovation: Technological, political and cultural. In R. Lehming & M. Kane (Eds.), *Improving schools: Using what we know* (pp. 17-41). Beverley Hills, CA: Sage.

House, E. R., & McQuillan, P. J. (1998). Three perspectives on school reform. In A. Hargreaves, A. Lieberman, M. Fullan & D. Hopkins (Eds.), *International handbook of educational change* (pp. 198-213). Dordrecht, the Netherlands: Kluwer Academic.

Howe, K. R. (1988). Against the quantitative-qualitative incompatibility thesis or dogmas die hard. *Educational Researcher, 17*(8), 10-16.

Howe, K. R. (1992). Getting over the quantitative-qualitative debate. *American Journal of Education, 100*(2), 236-256.

Huff, S. (2009). Build, promote, guide, provide, monitor: Action words for principals as instructional leaders in assessment. In T. R. Guskey (Ed.), *The principal as assessment leader* (pp. 31-51). Bloomington, IN: Solution Tree Press.

Hume, A., & Coll, R. K. (2009). Assessment of learning, for learning, and as learning: New Zealand case studies. *Assessment in Education: Principles, Policy & Practice, 16*(3), 269-290.

Hutchinson, C., & Hayward, L. (2005). The journey do far: Assessment for learning in Scotland. *The Curriculum Journal, 16*(2), 225-248.

Jakicic, C. (2009). A principal's guide to assessment. In T. R. Guskey (Ed.), *The principal as assessment leader* (pp. 53-70). Bloomington, IN: Solution Tree Press.

James, M. (2006). Assessment, teaching and theories of learning. In J. Gardner (Ed.), *Assessment and learning* (pp. 47-60). London: Sage.

James, M. (2008). Assessment and learning. In S. Swaffield (Ed.), *Unlocking assessment: Understanding for reflection and application* (pp. 20-35). Oxon, OX: Routledge.

James, M., & Lewis, J. (2012). Assessment in harmony with our understanding of learning: Problems and possibilities. In J. Gardner (Ed.), *Assessment and learning* (2nd ed., pp. 187-205). London: Sage.

James, M., & Pedder, D. (2006). Beyond method: Assessment and learning practices and values. *The Curriculum Journal, 17*(2), 109-138. doi: 10.1080/09585170600792712

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research, 1*(2), 112-133.

Joncas, M. (2008). TIMSS 2007 sample design. In M. O. Martin, I. V. S. Mullis & P. Foy (Eds.), *TIMSS 2007 international science report: Findings from IEA's Trends in International Mathematics and Science Study at the fourth and eighth grades* (pp. 77-92). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College, Lynch School of Education.

Jones, A., & Moreland, J. (2005). The importance of pedagogical content knowledge in assessment for learning practices: A case-study of a whole-school approach. *The Curriculum Journal, 16*(2), 193-206.

Kapambwe, W. M. (2010). The implementation of school based continous assessment (CA) in Zambia. *Educational Research and Reviews, 5*(3), 99-107.

Kay, K. (2010). Foreword. 21st century skills: Why they matter, what they are, and how we get there. In J. Bellenca & R. Brandt (Eds.), *21st century skills: Rethinking how students learng* (pp. xiii-xxxi). Bloomington, IN: Solution Tree.

Kelly, G. J. (2006). Epistemology and educational research. In J. L. Green, G. Camilli & P. B. Elmore (Eds.), *Handbook of complementary methods in education research*. Mahwah, N.J.: Lawrence Erlbaum Associates.

King, M. B., Schroeder, J., & Chawszczewski, D. (2001). Authentic assessment and student performance in inclusive schools. *Brief No. 5*. Retrieved from http://archive.wceruw.org/riser/Brief%205.pdf

Kirkup, C. (2006). Using assessment information to inform teaching and learning. *Education 3-13, 34*(2), 153-162.

Kirton, A., Hallam, S., Peffers, J., Robertson, P., & Stobart, G. (2007). Revolution, evolution or a trojan horse? Piloting assessment for learning in some Scottish primary schools. *British Educational Research Journal, 33*(4), 605-627.

Klenowski, V. (2009). Assessment for learning revisited: An Asian-Pacific perspective. *Assessment in Education: Principles, Policy & Practice, 16*(3), 263-268.
Kliebard, H. M. (2004). *The struggle for the American curriculum: 1893-1958* (3rd ed.). New York: RoutledgeFalmer.
Koh, K. H. (2011a). *Improving teachers' assessment literacy*. Singapore: Prentice Hall, Pearson.
Koh, K. H. (2011b). Improving teachers' assessment literacy through professional development. *Teaching Education, 22*(3), 255-276.
Koh, K. H., Lee, A. N., Gong, W., & Wong, H. M. (2006, 21-26 May 2006). *Development of the Singapore prototype classroom assessment tasks: Innovative tools for improving student learning and performance.* Paper presented at the 32nd Annual Conference: International Association for Educational Assessment (IAEA), Singapore.
Koh, K. H., Lee, A. N., Tan, W., Wong, H. M., Guo, L., Lim, T. M., . . . Tan, S. (2005). *Looking collaboratively at the quality of teachers' assessment tasks and student work in Singapore schools*. Paper presented at the 2005 Centre for Research in Pedagogy and Practice Conference: Redesigning pedagogy - Research, policy and practice, Singapore.
Koh, K. H., & Luke, A. (2009). Authentic and conventional assessments in Singapore schools: An empirical study of teacher assignments and student work. *Assessment in Education: Principles, Policy & Practice, 16*(3), 291-318.
Koh, k. H., Tan, C., & Ng, P. T. (2012). Creating thinking schools through authentic assessment: The case in Singapore. *Educational Assessment, Evaluation and Accountability, 24*(2), 135-149.
Kreber, C., Klampfleitner, M., McCune, V., Bayne, S., & Knottenbelt, M. (2007). What do you mean by "authentic"? A comparative review of the literature on conceptions of authenticity in teaching. *Adult Education Quarterly, 58*(1), 22-43.
Kvale, S. (1996). InterViews: An introduction to qualitative research interviewing. Thousand Oaks, CA: Sage.
Ladwig, J. G. (1998). Authentic school reform. *Discourse: Studies in the Cultural Politics of Education, 19*(1), 113-119.
Ladwig, J. G., Smith, M., Gore, J., Amosa, W., & Griffiths, T. (2007). *Quality of pedagogy and student achievement: Multi-level replication of authentic pedagogy*. Paper presented at the Australian Association for Research in Education Conference (25-29 November), Fremantle, Perth, Australia.
Lambert, D., & Lines, D. (2000). *Understanding assessment: Purposes, perceptions, practice*. London: RoutledgeFalmer.
Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge: Cambridge University Press.
Learning and Teaching Scotland. (2006). Assessment is for learning: Self-assessment toolkit. Retrieved March 20, 2010, from http://www.ltscotland.org.uk/assess/about/whatisaifl.asp
Leat, D., & Nichols, A. (2000). Brains on the table: Diagnostic and formative assessment through observation. *Assessment in Education: Principles, Policy & Practice, 7*(1), 103-121.
Lee, H. L. (2004, November 20, 2008). National Day Rally 2004 Speech, Sunday, 22 August 2004, at the University Cultural Centre, National University of Singapore. Retrieved 10 September, 2011, from http://www.getforme.com/pressreleases/leehl_220804_nationaldayrally2004.htm

Lee, H. L. (2012). Prime Minister Lee Hsien Loong's National Day Message 2012 (English). Retrieved 1 Septermber, 2012, from http://www.pmo.gov.sg/content/pmosite/mediacentre/speechesninterviews/primeminister/2012/August/national_day_message2012english.html

Lee, I. (2007). Feedback in Hong Kong secondary writing classrooms: Assessment for Learning or Assessment of Learning? *Assessing Writing, 12*(3), 180-198.

Lee, K. Y. (2000). *From third world to first: The Singapore story (1965-2000)*. New York: HarperCollins.

Lee, V. E., & Smith, J. B. (1994). High school restructuring and student achievement: A new study finds strong links (Issue Report No. 7). Madison, WI: Center on Organization and Restructuring of Schools.

Lee, V. E., & Smith, J. B. (1995). Effects of high school restructuring and size on early gains in achievement and engagement. *Sociology of Education, 68*(4), 241-270.

Lee, V. E., Smith, J. B., & Croninger, R. G. (1995). Another look at high school restructuring: More evidence that it improves student achievement and more insight into why. Madison, WI: Center on Organization and Restructuring of Schools.

Lee, V. E., Smith, J. B., & Croninger, R. G. (1997). How high school organization influences the equitable distribution of learning in mathematics and science. *Sociology of Education, 70*(2), 128-150.

Leighton, J. P., Gokiert, R. J., Cor, M. K., & Heffernan, C. (2010). Teacher beliefs about the cognitive diagnostic information of classroom versus large-scale tests: Implications for assessment literacy. *Assessment in Education: Principles, Policy & Practice, 17*(1), 7-21.

Leung, C., & Rea-Dickins, P. (2007). Teacher assessment as policy instrument: Contradictions and capacities. *Language Assessment Quarterly, 4*(1), 6-36.

Levin, B. (1994). Educational reform and the treatment of students in schools. *Journal of Educational Thought, 28*(1), 88-101.

Levin, B. (2000). Putting students at the center of education reform. *Journal of Educational Change, 1*(2), 155-172.

Liew, W. M. (2008). The realities of teaching amid the pressures of educational reform. In J. Tan & P. T. Ng (Eds.), *Thinking schools, learning nation: Contemporary issues and challenges* (pp. 104-134). Singapore: Pearson Education South Asia Pte Ltd.

Lim, E. P. Y., & Tan, A. (1999). Educational assessment in Singapore. *Assessment in Education: Principles, Policy & Practice, 6*(3), 391-404.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Newbury Park, CA: Sage.

Lingard, B., Mills, M., & Hayes, D. (2006). Enabling and aligning assessment for learning: Some research and policy lessons from Queensland. *International Studies in Sociology of Education, 16*(2), 83-103.

Looney, J. W. (2009). Assessment and innovation in education. OCED Working papers, Number 24, OECD Publishing. Retrieved 1 September, 2012, from http://dx.doi.org/10.1787/222814543073

Looney, J. W. (2011). Integrating formative and summative assessment: Progress toward a seamless system?, OECD Education Working Papers, No. 57, OECD Publishing. Retrieved 1 September, 2012, from http://dx.doi.org/10.1787/5kghx3kbl734-en

Luke, A., Freebody, P., Shun, L., & Gopinathan, S. (2005). Towards research-based innovation and reform: Singapore schooling in transition. *Asia Pacific Journal of Education, 25*(1), 5-28.

Luke, A., & Hogan, D. (2006). Redesigning what counts as evidence in educational policy: The Singapore model. In J. Ozga, T. Seddon & T. S. Popkewitz (Eds.), *The world year book of education* (pp. 170-184). New York: Routledge.

Luke, A., Matters, G., Hershell, P., Grace, N., Barrett, R., & Land, R. (2000). Draft New Basics project technical paper. Queensland, Australia.

Lund, T. (2005). The qualitative-quantitative distinction: Some comments. *Scandinavian Journal of Educational Research, 49*(2), 115-132.

Maclellan, E. (2004). Evidence of authentic achievement: The extent of disciplined enquiry in student teachers' essay scripts. *Australian Journal of Educational & Developmental Psychology, 4*, 71-85.

MacPhail, A., & Halbert, J. (2010). 'We had to do intelligent thinking during recent PE': Students' and teachers' experiences of assessment for learning in post-primary physical education. *Assessment in Education: Principles, Policy & Practice, 17*(1), 23-39. doi: 10.1080/09695940903565412

Madaus, G. F., & O'Dwyer, L. M. (1999). A short history of performance assessment: Lessons learned. *Phi Delta Kappan, 80*(9), 688-695.

Manning, C., Sisserson, K., Jolliffe, D., Buenrostro, P., & Jackson, W. (2008). Program evaluation as professional development: Building capacity for authentic intellectual achievement in Chicago small schools. *Education and Urban Society, 40*(6), 715-729.

Marks, H. M. (1995). Student engagement in the classrooms of restructuring schools. Madison, WI: Center on Organization and Restructuring of Schools.

Marsh, C. (2007). A critical analysis of the use of formative assessment in schools. *Educational Research for Policy and Practice, 6*(1), 25-29.

Marshall, B., & Drummond, M. J. (2006). How teachers engage with Assessment for Learning: Lessons from the classroom. *Research Papers in Education, 21*(2), 133-149.

Marshall, H. H. (1992a). Seeing, redefining, and supporting student learning. In H. H. Marshall (Ed.), *Redefining student learning* (pp. 1-32). Norwood, NJ: Ablex Publishing Corporation.

Marshall, H. H. (Ed.). (1992b). *Redefining student learning: Roots of educational change*. Norwoos, NJ: Ablex Publishing Corporation.

Martin, M. O., Mullis, I. V. S., & Foy, P. (2008). *TIMSS 2007 international science report: Findings from IEA's Trends in International Mathematics and Science Study at the fourth and eighth grades*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College, Lynch School of Education.

Martin, M. O., Mullis, I. V. S., Foy, P., & Stanco, G. M. (2012). *TIMSS 2011 international results in science*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and the International Association for the Evaluation of Educational Achievement.

Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., & Chrowstowski, S. J. (2004). *TIMSS 2003 International Science Report: Findings from the IEA's Trends in International Mathematics and Science Study at the fourth and eighth grades*. Chestnut Hill, MA: TIMSS and PIRLS International Study Center, Lynch School of Education, Boston College.

Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., Gregory, K. D., Smith, T. A., Chrostowski, S. J., . . . O'Connor, K. M. (2000). *TIMSS1999 international science report: Findings from IEA's Repeat of the Third International Mathematics and Science Study at the eighth*

grade. Chestnut Hill: TIMSS & PIRLS International Study Center, Boston, College, Lynch School of Education.
Martin, M. O., Mullis, I. V. S., Gregory, K. D., Hoyle, C., & Shen, C. (2000). *Effective schools in science and mathematics: IEA's Third International Mathematics and Science Study* (Vol. 2012). Chestnut Hill, MA: International Study Center, Lynch School of Education, Boston College.
Marton, F., Dall'Alba, G., & Tse, L. J. (1996). Memorizing and understanding. In D. A. Watkins & J. B. Biggs (Eds.), *The Chinese learner: Cultural, psychological and contextual influences* (pp. 69-83). Hong Kong: Comparative Education Research Centre and the Australian Council for Educational Research.
Matsumura, L. C., & Pascal, J. (2003). Teachers' assignments and student work: Opening a window on classroom practice (CSE Report 602). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles.
Maxwell, J. A. (2013). *Qualitative research design*. Thousand Oaks, CA: Sage Publications, Inc.
Mayer, D. P. (1999). Measuring instructional practice: Can policymakers trust survey data? *Educational Evaluation and Policy Analysis, 21*(1), 29-45.
McDonald, B., & Boud, D. (2003). The impact of self-assessment on achievement: The effects of self-assessment training on performance in external examinations. *Assessment in Education: Principles, Policy & Practice, 10*(2), 209-220.
McMillan, J. H. (2001). Secondary teachers' classroom assessment and grading practices. *Educational Measurement: Issues and Practice, 20*(1), 20-32.
McMillan, J. H., Myran, S., & Workman, D. (2002). Elementary teachers' classroom assessment and grading practices. *Journal of Educational Research, 95*(4), 203-213.
McNair, S., Bhargava, A., Adams, L., Edgerton, S., & Kypros, B. (2003). Teachers speak out on assessment practices. *Early Childhood Education Journal, 31*(1), 23-31.
McQuillan, P. J. (2005). Possibilities and pitfalls: A comparative analysis of student empowerment. *American Educational Research Journal, 42*(4), 639-670.
McQuillan, P. J., D'Souza, L. A., Scheopner, A. J., Miller, G. R., Gleeson, A. M., Mitchell, K., . . . Cochran-Smith, M. (2009). Reflecting on pupil learning to promote social justice: A catholic university's approach to assessment. *Catholic Education: A Journal of Inquiry and Practice, 13*(2), 157-184.
Meier, D. (1998). Authenticity and educational change. In A. Hargreaves, A. Lieberman, M. Fullan & D. Hopkins (Eds.), *International handbook of educational change* (pp. 596-615). Dordrecht, The Netherlands: Kluwer Academic Publishers.
Merriam-Webster. (2011). Merriam-Webster Online Dictionary, from http://www.merriam-webster.com/
Mertler, C. A. (1999). Assessing student performance: A descriptive study of the classroom assessment practices of Ohio teachers. *Education, 120*(2), 285-296.
Mertler, C. A. (2005). The Role of Classroom Experience in Preservice and Inservice Teachers' Assessment Literacy. *Mid-Western Educational Researcher, 18*(4; 4), 25-34.
Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Thousand Oaks, CA: Sage.
Miller, D., & Lavin, F. (2007). 'But now I feel I want to give it a try': Formative assessment, self-esteem and a sense of competence. *The Curriculum Journal, 18*(1), 3-25.

Milnes, T., & Cheng, L. (2008). Teachers' assessment of ESL students in mainstream classes: Challenges, strategies, and decision-making. *TESL Canada Journal, 25*(2), 49-64.
Ministry of Education. (1998). Ministry of Education's response to the External Curriculum Review report Retrieved 10 September, 2011, from http://www.moe.gov.sg/media/press/1998/980321.htm
Ministry of Education. (2009). *Report of the Primary Education Review and Implementation (PERI) Committee* Singapore: Ministry of Education Retrieved from http://planipolis.iiep.unesco.org/upload/Singapore/Singapore_PERI_2009.pdf.
Ministry of Education. (2012). International studies affirm Singapore students' strengths in reading, mathematics and science. Retrieved 11 December, 2012, from http://www.moe.gov.sg/media/press/2012/12/international-studies-affirm-s.php
Mitra, D. L. (2009). Student voice and student roles in educational policy and policy reform. In G. Sykes, B. Schneider, D. N. Plank & T. G. Ford (Eds.), *Handbook of education policy research* (pp. 819-830). New York: American Educational Research Association & Routledge.
MOE [Bluesky]. (2005, 20 May 2009). Teach Less, Learn More. Retrieved 20 April, 2009, from http://www3.moe.edu.sg/bluesky/print_tllm.htm
Mourshed, M., Chijioke, C., & Barber, M. (2010). *How the world's most improved school systems keep getting better*. London: McKinsey & Company.
Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.
Mullis, I. V. S., Martin, M. O., Smith, T. A., Garden, R. A., Gregory, K. D., Gonzalez, E. J., . . . O'Connor, K. M. (2003). *TIMSS: Assessment frameworks and specifications 2003* (2nd ed.). Chestnut Hill, MA: TIMSS International Study Center, Lynch School of Education, Boston College.
Natriello, G. (1987). The impact of evaluation processes on students. *Educational Psychologist, 21*(2), 155-175.
Newmann, F. M. (1992). Higher-order thinking and prospects for classroom thoughtfulness. In F. M. Newmann (Ed.), *Student engagement and achievement in American secondary schools* (pp. 62-91). New York: Teachers College Press.
Newmann, F. M. (1996). Center on Organization and Restructuring of Schools: Activities and accomplishments, 1990-1996. Final Report. (pp. 1-25). Madison, WI: Center on Organization and Restructuring of Schools & Wisconsion Center for Education Research, Madison.
Newmann, F. M., & Archbald, D. A. (1992). The nature of authentic academic achievement. In H. Berlak, F. M. Newmann, E. Adams, D. A. Archbald, T. Burgess, J. Raven & T. A. Romberg (Eds.), *Towards a new science of educational testing and assessment* (pp. 71-83). Albany, NY: State University of New York Press.
Newmann, F. M., & Associates. (1996). *Authentic achievement: Restructuring schools for intellectual quality*. San Francisco, CA: Jossey-Bass Publishers.
Newmann, F. M., Brandt, R., & Wiggins, G. (1998). An exchange of views on "Semantics, psychometrics, and assessment reform: A close look at "authentic" assessments". *Educational Researcher, 27*(6), 19-22.

Newmann, F. M., Bryk, A. S., & Nagaoka, J. K. (2001). *Authentic intellectual work and standardized tests: Conflict or coexistence*. Chicago, IL: Consortium on Chicago School Research.

Newmann, F. M., King, M. B., & Carmichael, D. L. (2007). Authentic instruction and assessment: Common standards for rigor and relevance in teaching academic subjects. Des Moines, IA: Department of Education, Iowa.

Newmann, F. M., Lopez, G., & Bryk, A. S. (1998). The quality of intellectual work in Chicago Schools: A baseline report. Chicago, IL: Consortium on Chicago School Research.

Newmann, F. M., Marks, H. M., & Gamoran, A. (1996). Authentic pedagogy and student performance. *American Journal of Education, 104*(4), 280-312.

Newmann, F. M., Secada, W. G., & Wehlage, G. G. (1995). *A guide to authentic instruction and assessment: Vision, standards and scoring*. Madison, WI: Wisconsin Center for Education Research.

Newmann, F. M., Smith, B., Allensworth, E., & Bryk, A. S. (2001). Instructional program coherence: What it is and why it should guide school improvement policy. *Educational Evaluation and Policy Analysis, 23*(4), 297-321.

Newmann, F. M., & Wehlage, G. G. (1993). Five standards of authentic instruction. *Educational Leadership, 50*(7), 8-12.

Newmann, F. M., & Wehlage, G. G. (1995). Successful school restructuring: A report to the public and educators by the Center on Organization and Restructuring of Schools. Wisconsin, MI: Wisconsin Center for Educational Research.

Newton, P. E. (2007). Clarifying the purposes of educational assessment. *Assessment in Education: Principles, Policy & Practice, 14*(2), 149-170.

Ng, P. T. (2008). Thinking schools, learning nation. In J. Tan & P. T. Ng (Eds.), *Thinking Schools, Learning Nation: Contemporary Issues and Challenges* (pp. 1-6). Singapore: Pearson Education South Asia Ptd Ltd.

Nitko, A. J., & Brookhart, S. M. (2011). *Educational assessment of students* (6th ed.). Boston, MA: Pearson Education, Inc.

Norris, S. P., & Ennis, R. H. (1989). *Evaluating critical thinking*. Pacific Grove, CA: Critical Thinking Press & Software.

North Central Regional Educational Laboratory & the Metiri Group. (2003). *enGauge 21st century skills: Literacy in the digital age*. Chicago, IL: North Central Regional Educational Library.

Nouršis, M. J. (2008). *SPSS16.0 Guide to data analysis*. Upper Saddle River, NJ: Prentice Hall.

OECD. (2009). PISA 2009 Assessment Framework – Key Competencies in reading, mathematics and science Retrieved from http://www.oecd.org/document/44/0,3746,en_2649_35845621_44455276_1_1_1_1,00.html

OECD. (2011). Lessons from PISA for the United States, Strong Performers and Successful Reformers in Education, OCED Publishing. Retrieved from doi:http://dx.doi.org/10.1787/9789264096660-en

Ohlsen, M. T. (2007). Classroom assessment practices of secondary school members of NCTM. *American Secondary Education, 36*(1), 4-14.

Olson, J. F., Martin, M. O., & Mullis, I. V. S. (2008). TIMSS 2007 Technical Report. Chestnut Hill, M.A.: TIMSS & PIRLS International Study Center, Boston College, Lynch School of Education.

Onewuegbuzie, A. J., & Teddlie, C. (2003). A framework for analyzing data in mixed methods research. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 351-383). Thousand Oaks, CA: Sage.

Online Etymology Dictionary. (2013). Online etymology dictionary. Retrieved 1 September, 2013, from http://www.etymonline.com/index.php?allowed_in_frame=0&search=authentic&searchmode=none

Onwuegbuzie, A. J., & Combs, J. P. (2010). Emergent data analysis techniques in mixed methods research In A. Tashakkori & C. Teddlie (Eds.), *Sage handbook of mixed methods in social and behavioral research* (2nd ed., pp. 397-430). Thousand Oaks, CA: Sage Publications.

Onwuegbuzie, A. J., & Johnson, R. B. (2006). The validity issue in mixed research. *Research in the Schools, 13*(1), 48-63.

Organisation for Economic Co-operation and Develoment. (2005). *The definition and selection of key competencies: Executive summary*. Paris: Author.

Paige, R. R., & Witty, E. P. (2008). Assessment as instructional support. In C. A. Dwyer (Ed.), *The future of assessment: Shapging teaching and learning* (pp. 215-228). New York: Lawrence Erlbaum Associates.

Palm, T. (2008). Performance assessment and authentic assessment: A conceptual analysis of the literature. *Practical Assessment, Research and Evaluation, 13*(4), 1-11. Retrieved from http://pareonline.net/getvn.asp?v=13&n=4

Parr, J. M., & Timperley, H. S. (2010). Feedback to writing, assessment for teaching and learning and student progress. *Assessing writing, 15*(2), 68-85.

Partnership for 21st Century Skills. (2011a, 2011). About us: Our history Retrieved 6 November, 2011, from http://www.p21.org/about-us/our-history

Partnership for 21st Century Skills. (2011b). P21 framework definitions. Retrieved 6 November, 2011, from http://www.p21.org/storage/documents/P21_Framework_Definitions.pdf

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academic Press.

Pellegrino, J. W., & Goldman, S. R. (2008). Beyond rhetoric: Realities and complexities of integrating assessment into classroom teaching and learning. In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 7-52). New York: Erlbaum Associates.

Perkins, D. (1999). The many faces of constructivism. *Educational Leadership, November*, 6-11.

Plake, B. S., & Impara, J. C. (1997). Teacher assessment literacy: What do teachers know about assessment? In G. D. Phye (Ed.), *Handbook of classroom assessment: Learning, achievement and adjustment* (pp. 53-68). San Diego, CA: Academic Press.

Popham, W. J. (2008). *Transformative assessment*. Alexandria, VA: ASCD.

Priestley, M. (2005). Making the most of the Curriculum Review: Some reflections on supporting and sustaining change in schools. *Scottish Educational Review, 37*(1), 29-38.

Priestley, M., & Sime, D. (2005). Formative assessment for all: A whole-school approach to pedagogic change. *The Curriculum Journal, 16*(4), 475-492.

Pryor, J., & Crossouard, B. (2008). A socio-cultural theorisation of formative assessment. *Oxford Review of Education, 34*(1), 1-20. doi: 10.1080/03054980701476386

Pryor, J., & Torrance, H. (1998). Formative assessment in the classroom: Where psychological theory meets social practice. *Social Psychology of Education, 2*(2), 151-176.
Purcell-Gates, V., Duke, N. K., & Martineau, J. A. (2007). Learning to read and write genre-specific text: Roles of authentic experience and explicit teaching. *Reading Research Quarterly, 42*(1), 8-45.
QSRLS. (2001). Submitted to Education Queensland by the School of education, University of Queensland. State of Queensland (Department of Education), Brisbane.
Rahm, J., Miller, H. C., Hartley, L., & Moore, J. C. (2003). The value of an emergent notion of authenticity: Examples from two student/teacher-scientist partnership programs. *Journal of Research in Science Teaching, 40*(8), 737-756.
Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science, 28*(1), 4-13.
Rea-Dickins, P., & Gardner, S. (2000). Snares and silver bullets: Disentangling the construct of formative assessment. *Language Testing, 17*(2), 215-243.
Read, A., & Hurford, D. (2010). 'I know how to read longer novels' - developing pupils' success criteria in the classroom. *Education 3-13, 38*(1), 87-100.
Remesal, A. (2007). Educational reform and primary and secondary teachers' conceptions of assessment: The Spanish instance, building upon Black and Wiliam (2005). *The Curriculum Journal, 18*(1), 27-38.
Remesal, A. (2011). Primary and secondary teachers' conceptions of assessment: A qualitative study. *Teaching and Teacher Education, 27*(2), 472-482.
Research Institute on Secondary Education Reform (RISER). (2001). *Standards and scoring criteria for assessment tasks and student performance*. Madison, WI: Author.
Resnick, L. B. (1987a). The 1987 Presidential Address: Learning in school and out. *Educational Researcher, 16*(9), 13-20+54.
Resnick, L. B. (1987b). *Education and learning to think*. Washington, DC: National Academy Press.
Resnick, L. B. (1989). Introduction. In L. B. Resnick (Ed.), *Knowing, learning and instruction: Essays in honor of Robert Glaser* (pp. 1-24). Hillsdale, NJ: Lawrence Erlbaum Associates.
Rezulli, J. S., Gentry, M., & Reis, S. M. (2004). A time and a place for authentic learning. *Educational Leadership, 62*(1), 73-77.
Richards, C. (2004). From old to new learning: Global imperatives, exemplary Asian dilemmas and ICT as a key to cultural change in education. *Globalisation, Societies and Education, 2*(3), 337-353.
Richardson, V. (1997). Constructivist teaching and teacher education: Theory and practice. In V. Richardson (Ed.), *Constructivist teacher education: Building new understandings* (pp. 3-14). London: The Falmer Press.
Riggan, M., & Oláh, L. N. (2011). Locating interim assessments within teachers' assessment practice. *Educational Assessment, 16*(1), 1-14.
Robinson, K. (2010). Changing paradigms [from The Royal Society for the Encouragement of Arts, Manufactures and Commerce (RSA)] Retrieved 25 October, 2011, from http://www.youtube.com/watch?v=zDZFcDGpL4U
Roos, B., & Hamilton, D. (2005). Formative assessment: A cybernetic viewpoint. *Assessment in Education: Principles, Policy & Practice, 12*(1), 7-20.
Rudduck, J. (1991). *Innovation and change: Developing involvement and understanding*. Buckingham: Open University Press.

Rudduck, J. (2002). The 2002 SERA lecture: The transformative potential of consulting young people about teaching, learning, and schooling. *Scottish Educational Review, 34*(2), 133-137.
Rudduck, J. (2007). Student voice, student engagement, and school reform. In D. Thiessen & A. Cook-Sather (Eds.), *International Handbook of Student Experience in Elementary and Secondary School* (pp. 587-610). Dordrecht, The Netherlands: Springer.
Rudduck, J., & Flutter, J. (2004). *How to improve your school: Giving pupils a voice*. London: Continuum Books.
Ruiz-Primo, M. A., & Furtak, E. M. (2006a). Exploring teachers' informative formative assessment practices and students' understanding in the context of scientific inquiry. *Journal of Research in Science Teaching, 44*(1), 57-84. doi: 10.1002/tea.20163
Ruiz-Primo, M. A., & Furtak, E. M. (2006b). Informal formative assessment and scientific inquiry: Exploring teachers' practices and student learning. *Educational Assessment, 11*(3-4), 237-263.
Ruiz-Primo, M. A., & Furtak, E. M. (2007). Exploring Teachers' Informal Formative Assessment Practices and Students' Understanding in the Context of Scientific Inquiry. *Journal of Research in Science Teaching, 44*(1), 57-84.
Ryan, G. W., & Bernard, H. R. (2000). Data management and analysis methods. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 769-802). Thousand Oaks, CA: Sage.
Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*, 119-144.
Sadler, D. R. (1998). Formative Assessment: Revisiting the territory. *Assessment in Education: Principles, Policy & Practice, 5*(1), 77-84.
Sandelowski, M. (2001). Real qualitative researchers do not count: The use of numbers in qualitative research. *Research in Nursing & Health, 24*(3), 230-240.
Sandelowski, M., Voils, C. I., & Knafl, G. (2009). On quantitizing. *Journal of Mixed Methods Research, 3*(3), 208-222.
Schiro, M. S. (2008). *Curriculum theory*. Los Angeles: Sage.
Schleicher, A. (2010). International comparisons of student learning outcomes. In A. Hargreaves, A. Lieberman, M. Fullan & D. Hopkins (Eds.), *Second international handbook of educational change* (pp. 485-504). Dordrecht, The Netherlands: Springer.
Scriven, M. (1967). The methodology of evaluation. In R. Tyler, R. Gagne & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (pp. 39-83). Chicago, IL: Rand McNally & Company.
Segers, M., & Tillema, H. (2011). How do Dutch secondary teachers and students conceive the purpose of assessment? *Studies in Educational Evaluation, 37*(1), 49-54.
Sfard, A. (1998). On Two Metaphors for Learning and the Dangers of Choosing Just One. *Educational Researcher, 27*(2), 4-13.
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Belmont, CA: Wadsworth.
Shanmugaratnam, T. (2004). To light a fire: Enabling teachers, nurturing students. Speech at the MOE Workplan Seminar on Wednesday, 29 September 2004. Retrieved 4 April 2009, 2009, from http:www.moe.gov.sg/media/speeches/2004/sp20040929_print.htm
Shanmugaratnam, T. (2005a, 2 January 2008). Achieving quality: Bottom up initiative, top down support. Speech by Mr Tharman Shanmugaratnam, Minister for Education, at the MOE

Work Plan Seminar 2005, on Thursday, 22 September 2005 at 10.00am at the Ngee Ann Polytechnic Convention Centre Retrieved 1 September, 2012, from http://www.moe.gov.sg/media/speeches/2005/sp20050922.htm

Shanmugaratnam, T. (2005b, 2 January 2008). Achieving quality: Bottom up initiative, top down support. Speech by Mr Tharman Shanmugaratnam, Minister for Education, at the MOE Work Plan Seminar 2005, on Thursday, 22 September 2005. Retrieved 20 November, 2008, from http://www.moe.gov.sg/media/speeches/2005/sp20050922.htm

Shanmugaratnam, T. (2007). Having every child succeed. Speech at the MOE Workplan Seminar 2007 on Tuesday, 2 October 2007. Retrieved 4 April 2009, 2009, from http://moe.gov.sg/media/speeches/2007/sp20071002_print.htm

Shavelson, R. J. (1996). *Statistical reasoning for the behaviorial sciences* (3rd ed.). Boston, MA: Allyn and bacon.

Shavelson, R. J., Young, D. B., Ayala, C. C., Brandon, P. R., Furtak, E. M., Ruiz-Primo, M. A., . . . Yin, Y. (2008). On the impact of curriculum-embedded formative assessment on learning: A collaboration between curriculum and assessment developers. *Applied Measurement in Education, 21*(4), 295-314.

Shepard, L. A. (1989). Why we need better assessments. *Educational Leadership, April*, 4-9.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4-14.

Shepard, L. A. (2001). The role of classroom assessment in teaching and learning. In V. Richardson (Ed.), *Handbook of Research on Teaching* (4th ed., pp. 1066-1101). Washington D.C.: AERA.

Shepard, L. A. (2006). Classroom assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 623-646). Westport, CT: Praeger

Shepard, L. A., & Kirst, M. W. (1991). Interview on assessment issues with Lorrie Shepard. *Educational Researcher, 20*(2), 21-23+27.

Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*(1), 1-22.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153-189.

Smith, E., & Gorard, S. (2005). 'They don't give us our marks': The role of formative feedback in student progress. *Assessment in Education: Principles, Policy & Practice, 12*(1), 21-38.

Smith, J. R. (1928). *Geography and our need of it*. Chicago: American Library Association.

Smith, M. L. (2006). Multiple methodology in educational research. In J. L. Green, G. Camilli & P. B. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 457-475). Mahwah, NJ: American Educational Research Association.

Smith, S., Layng, J., & Jones, M. (1995). *The 3-D View of Authentic Assessment (ED391486).* Paper presented at the Annual Conference of the International Visual Literacy Association, Chicago, IL. http://www.eric.ed.gov/PDFS/ED391486.pdf

Snape, P., & Fox-Turnbull, W. (2011). Perspectives of authenticity: implementation in technology education. *International Journal of Technology and Design Education*, 1-18. doi: 10.1007/s10798-011-9168-2

Splitter, L. J. (2009). Authenticity and constructivism in education. *Studies in Philosophy and Education, 28*(2), 135-151.

Stemler, S. E., & Tsai, J. (2008). Best practices in interrater reliability: Three common approaches. In J. W. Osbourne (Ed.), *Best practices in quantitative methods* (pp. 29-49). Los Angeles, CA: Sage.
Stiggins, R. J. (1991). Assessment literacy. *Phi Delta Kappan, 72*(7), 534-539.
Stiggins, R. J. (1992). High quality classroom assessment: What does it really mean? *Educational Measurement: Issues and Practice, 11*(2), 35-39.
Stiggins, R. J. (2001). The principal's leadership role in assessment. *NASSP Bulletin, 85*(621), 13-26.
Stiggins, R. J. (2002). Assessment crisis: The absence of assesment FOR learning. *Phi Delta Kappan, 83*(10), 758-765.
Stiggins, R. J., & Bridgeford, N. J. (1985). The ecology of classroom assessment. *Journal of Educational Measurement, 22*(4), 271-286.
Stobart, G. (2008). *Testing times: The uses and abuses of assessment*. London: Routledge.
Strike, K. A. (2006). The ethics of educational research. In J. L. Green, G. Camilli & P. B. Elmore (Eds.), *Handbook of complementary methods in education research*. Mahwah, N.J.: Lawrence Erlbaum Associates.
Suah, S. L., & Ong, S. L. (2012). Investigating assessment practices of in-service teachers. *International Online Journal of Educational Sciences, 4*(1), 91-106.
Suurtamm, C., Koch, M., & Arden, A. (2010). Teachers' assessment practices in mathematics: Classrooms in the context of reform. *Assessment in Education: Principles, Policy & Practice, 17*(4), 399-417.
Tan, K. (2008). Rethinking TLLM and its consequantial effects on assessment. In J. Tan & P. T. Ng (Eds.), *Thinking schools, learning nation: Contemporary issues and challenges* (pp. 246-257). Singapore: Pearson Education South Asia Pte Ltd.
Tan, S. Y. (2000). Sharing a vision, nationwide: The Thinking Schools, Learning Nation initiative of Singapore. In P. Senge, N. Cambron-McCabe, T. Lucas, B. Smith, J. Dutton & A. Kleiner (Eds.), *Schools that learn: A fifth discipline fieldbook for educators, parents, and everyone who cares about education* (pp. 483-488). New York, NY: Doubleday.
Tan, Y. K., Chow, H. K., & Goh, C. (2008). *Examinations in Singapore: Change and continuity (1891-2007)*. Singapore: World Scientific.
Tanner, D. E. (2001). Authentic assessments: A solution, or part of the problem? *The High School Journal, 85*(1), 24-29.
Tashakkori, A., & Teddlie, C. (1998). *Mixed methodology: Combining qualitative and quantitative approaches* (Vol. 46). Thousand Oaks, CA: Sage.
Tashakkori, A., & Teddlie, C. (2008). Quality of inferences in mixed methods research: Calling for an integrative framework. In M. M. Bergman (Ed.), *Advances in mixed methods research* (pp. 101-119). Thousand Oaks, CA: Sage.
Teddlie, C., & Tashakkori, A. (2003). Major issues and controversies in the use of mixed methods in the social and behavioral sciences. In A. Takshakkori & C. Teddlie (Eds.), *Handbook of Mixed Methods in Social and Behavioral Research* (pp. 3-50). Thousand Oaks, California: Sage.
Teo, C. H. (1998). Talking points for RADM Teo Chee Hean, Minister for Education and 2nd Minister for Defence, at the 1998 World Economic Forum Annual Meeting "Educating tomorrow's global citizen: What should we be teaching today's youth?" (Tuesday, 3 Feb 1998) Retrieved 25 September, 2011, from http://www.moe.gov.sg/media/speeches/1998/030298.htm

Teo, C. H. (2002). Opening speech by RADM Teo Chee Hean, Minister for Education and Second Minister for Defence on the JC/Upper Secondary Review Committee recommendations at parliament on 25 Nov 2002 Retrieved 26 September, 2011, from http://www.moe.gov.sg/media/speeches/2002/sp27112002.htm

Terwilliger, J. S. (1997). Semantics, psychomterics, and assesment reform: A close look at "authentic" assessments. *Educational Researcher, 26*(8), 24-27.

Terwilliger, J. S. (1998). Rejoinder: Response to Wiggins and Newmann. *Educational Researcher, 27*(6), 22-23.

Thompson, M., & Wiliam, D. (2008). Tight but loose: A conceptual framework for scaling up school reform. In E. C. Wylie (Ed.), *Tight but loose: Scaling up teacher professional development in diverse contexts*. Princeton, NJ: Educational Testing Service.

Thorndike, R. M., & Thorndike-Christ, T. (2010). *Meausrement and evaluation in psychology and education* (8th ed.). Boston, MA: Pearson.

Tierney, R. D. (2006). Changing practices: Influences on classroom assessment. *Assessment in Education: Principles, Policy & Practice, 13*(3), 239-264.

Tiknaz, Y., & Sutton, A. (2006). Exploring the role of assessment tasks to promote formative assessment in Key Stage 3 geography: Evidence from twelve teachers. *Assessment in Education: Principles, Policy & Practice, 13*(3), 327-343.

Tinzmann, M., Jones, B. F., & Pierce, J. (1991). Changing societal needs: Changing how we think about curriculum and instruction. In C. Collins & H. N. Mangieri (Eds.), *Teaching thinking: An agenda for the 21st century* (pp. 185-220). Hillsdate, NJ: Lawrence Erlbaum Associates.

Torrance, H. (2007). Assessment as learning? How the use of explicit learning objectives, assessment criteria and feedback in post-secondary education and training can come to dominate learning. *Assessment in Education: Principles, Policy & Practice, 14*(3), 281-294.

Torrance, H. (Ed.). (1995). *Evaluating authentic assessment: Problems and possibilities in new approaches to assessment*. Buckingham, UK: Open University Press.

Torrance, H., & Pryor, J. (1998). *Investigating formative assessment : Teaching, learning and assessment in the classroom*. Buckingham: Open University Press.

Torrance, H., & Pryor, J. (2001). Developing formative assessment in the classroom: Using action research to explore and modify theory. *British Educational Research Journal, 27*(5), 615-631.

Trepanier-Street, M. L., McNair, S., & Donegan, M. M. (2001). The views of teachers on assessment: A comparison of lower and upper elementary teachers. *Childhood Education International, 15*(2), 234-241.

Trilling, B., & Fadel, C. (2009). *21st century skills: Learning for life in our times*. San Francisco, CA: Jossey-Bass.

Tyack, D., & Cuban, L. (1995). *Tinkering toward Utopia: A century of public school reform.* Cambridge, MA: Harvard University Press.

Tyack, D., & Tobin, W. (1994). The grammar of schooling: Why is it so hard to change. *American Educational Research Journal, 31*(3), 451-480.

Vagle, N. M. (2009). Finding meaning in the numbers. In T. R. Guskey (Ed.), *The principal as assessment leader* (pp. 149-173). Bloomington, IN: Solution Tree Press.

van't Hooft, M. (2005). The effect of the "Ohio schools going solar" project on student perceptions of the quality of learnig in middle school science. *Journal of Research on Technology in Education, 37*(3), 221-243.
Viadero, D. (2005). 'Mixed methods' research examined. *Education Week, 24*(20), 1,20. Retrieved from http://www.apa.org/ed/schools/cpse/publications/mix-methods.pdf
Volante, L., & Beckett, D. (2011). Formative assessment and the contemporary classroom: Synergies and tensions between research and practice. *Canadian Journal of Education, 34*(2), 239-255.
Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological process*. Cambridge, MA: Harvard University Press.
Wagner, T. (2008). *The global achievement gap: Why even our best schools don't teach the new survival skills our children need - and what we can do about it*. New York: Basic Books.
Watkins, C. (2011). Learning: A sense-maker's guide. Retrieved 20 November, 2011, from http://www.atl.org.uk/Images/Learning%20a%20sense%20makers%20guide%20-%202011.pdf
Watkins, D. A., & Biggs, J. B. (Eds.). (1996). *The Chinese learner: Cultural, psychological and contextual influences*. Hong Kong: Comparative Education Research Centre and the Australian Council for Educational Research.
Watkins, D. A., & Biggs, J. B. (Eds.). (2001). *Teaching the Chinese learner: Psychological and pedagogical perspectives*. Hong Kong: Comparative Education Research Centre, The University of Hong Kong.
Webb, M., & Jones, J. (2009). Exploring tensions in developing assessment for learning. *Assessment in Education: Principles, Policy & Practice, 16*(2), 165-184.
Weeden, P., Winter, J., & Broadfoot, P. (2002). *Assessment: What's in it for schools?* Oxon, UK: RoutledgeFalmer.
Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70*(9), 703-713.
Wiggins, G. (1990). The case for authentic assessment. Retrieved from http://eric.ed.gov/PDFS/ED328611.pdf
Wiggins, G. (1992). Creating tests worth taking. *Educational Leadership, May*, 26-33.
Wiggins, G. (1993). Assessment: Authenticity, context, and validity. *Phi Delta Kappan, 75*(3), 200-208+210-214.
Wiggins, G., & McTighe, J. (2005). *Understanding by design* (Expanded 2nd ed.). Alexandria, VA: Association for Supervision and Curriculum Development.
Wiliam, D. (2010). An integrative summary of the research literature and implications for a new theory of formative assessment. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 18-40). New York: Routledge.
Wiliam, D., & Black, P. (1996). Meanings and consequences: A basis for distinguishing formative and summative functions of assessment? *British Educational Research Journal, 22*(5), 537-548.
Wiliam, D., Lee, C., Harrison, C., & Black, P. (2004). Teachers developing assessment for learning: Impact on student achievement. *Assessment in Education: Principles, Policy & Practice, 11*(1), 49-65.
Wiliam, D., & Thompson, M. (2008). Integrating assessment with learning: What will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 53-82). New York: Lawrence Erlbaum Associates.

Williams, J. A. (2010). 'You know what you've done right and what you've done wrong and what you need to improve on': New Zealand students' perspectives on feedback. *Assessment in Education: Principles, Policy & Practice, 17*(3), 301-314.

Wolf, D., Bixby, J., Glenn, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. *Review of Research in Education, 17*, 31-74.

Wong, M. W. (2007). Assessment for learning and teacher develoment: The experience of three Hong Kong teachers. *Teacher Development, 11*(3), 25-312. doi: 10.1080.13664530701644599

Yin, Y., Shavelson, R. J., Ayala, C. C., Ruiz-Primo, M. A., Brandon, P. R., Furtak, E. M., . . . Young, D. B. (2008). On the impact of formative assessment on student motivation, achievement, and conceptual change. *Applied Measurement in Education, 21*(4), 335-359.

Yip, J. S. K., Eng, S. P., & Yap, J. Y. C. (1994). 25 years of educational reform. In J. S. K. Yip & W. K. Sim (Eds.), *Evolution of educational excellence: 25 years of education in the Republic of Singapore* (pp. 1-32). Singapore: Longman.

Yu, K., & Frempong, G. (2012). Standardize and individualize - an unsolvable tension in assessment? *Education as Change, 16*(1), 143-157.

Zhang, Z., & Burry-Stock, J. A. (2003a). Classroom Assessment Practices and Teachers' Self-Perceived Assessment Skills. *Applied Measurement in Education, 16*(4), 323-342.

Zhang, Z., & Burry-Stock, J. A. (Writers). (2003b). Classroom assessment practices and teachers' self-perceived assessment skills, *Applied Measurement in Education*. US: Lawrence Erlbaum.

Appendix 1: Items on Teacher Assessment in TIMSS Cycles

TIMSS 1995 Science Teacher Questionnaire Main Survey	
Item 22: In assessing the work of the students in your science class, how much weight do you give each of the following types of assessment? a. Standardized tests produced outside the school b. Teacher-made short answer or essay tests that require students to describe or explain their reasoning c. Teacher made multiple choice, true-false and matching tests d. How well students do on homework assignments e. How well students do on projects or practical/laboratory exercises f. Observations of students g. Responses of students in class	The responses are in categories • None • Little • Quite a lot • A great deal
Item 23: How often do you use the assessment information you gather from students to… a. Provide students' grades or marks? b. Provide feedback to students? c. Diagnose students' learning problems? d. Report to parents? e. Assign students to different programs or tacks? f. Plan for future lessons?	The responses are in categories • None • Little • Quite a lot • A great deal
TIMSS 1999 Science Teacher Questionnaire Main Survey	
Item 19: In assessing the work of the students in your science class, how much weight do you give each of the following types of assessment? a. Standardized tests produced outside the school b. Teacher-made short answer or essay tests that require students to describe or explain their reasoning c. Teacher made multiple choice, true-false and matching tests d. How well students do on homework assignments e. How well students do on projects or practical/laboratory	The responses are based on frequency. The categories are • None • Little • Quite a lot • A great deal

exercises f. Observations of students g. Responses of students in class	
Item 20: How often do you use the assessment information you gather from students to… a. Provide students' grades or marks? b. Provide feedback to students? c. Diagnose students' learning problems? d. Report to parents? e. Assign students to different programs or tracks? Plan for future lessons?	The responses are based on frequency. The categories are • None • Little • Quite a lot • A great deal
TIMSS 2003 Teacher Questionnaire (Science Grade 8)	
Item 32: How often do you give a science test or examination to the <TIMSS class>? • About once a week • About every two weeks • About once a month • A few times a year • Never	Respondents shade the appropriate response.
Item 33: What item formats do you typically use in your science tests or examinations? • Only constructed-response • Mostly constructed response • About half-constructed response and half objective (E.g., multiple choice) • Mostly objective • Only objective	Respondents shade the appropriate response.
Item 34: How often do you include the following types of questions in your science tests or examinations? a. Questions requiring understanding of concepts, relationships, and processes b. Questions involving hypotheses and conclusions	Responses are in terms of frequency. The categories are: • Never or almost never • Sometimes • Always or almost always

c. Questions based on recall of facts or procedures	
TIMSS 2007 Teacher Questionnaire (Science Grade 8)	
Item 28: How much emphasis do you place on the following sources to monitor students' progress in science? a. Classroom tests (for example, teacher made or textbook tests) b. National or regional achievement tests c. Your professional judgment	The responses are in terms of the following categories: • No emphasis • Little emphasis • Some emphasis • Major emphasis
Item 29: How often do you give a science test or examination to the <TIMSS> class? • About once a week • About every two weeks • About once a month • A few times a year • Never	Respondents shade the appropriate response.
Item 30: What item formats do you typically use in your science tests or examinations? • Only constructed-response • Mostly constructed response • About half-constructed response and half objective (E.g., multiple choice) • Mostly objective • Only objective	Respondents shade the appropriate response.
Item 31: How often do you include the following types of questions in your science tests or examinations? a. Questions based on knowing facts and concepts b. Questions based on the application of knowledge and understanding c. Questions involving developing hypotheses and designing scientific investigations d. Questions requiring explanations or justifications	Responses are in terms of frequency. The categories are: • Never or almost never • Sometimes • Always or almost always

TIMSS 20011 Teacher Questionnaire (Science Grade 8)	
Item 25: How much emphasis do you place on the following sources to monitor students' progress in science? a. Evaluation of students' ongoing work b. Classroom tests (for example, teacher-made or textbook tests) c. National or regional tests	Responses are in terms of the degree of emphasis. The categories are: • Little or no emphasis • Some emphasis • Major emphasis
Item 26: How often do you give a science test or examination to this class? • About once a week • About every two weeks • About once a month • A few times a year • Never	Respondents check the appropriate response.
Item 27: How often do you include the following types of questions in your science tests or examinations? a. Questions based on knowing facts and concepts b. Questions based on the application of knowledge and understanding c. Questions involving developing hypotheses and designing scientific investigations d. Questions requiring explanations or justifications	Responses are in terms of frequency. The categories are: • Never or almost never • Sometimes • Always or almost always

Appendix 2: "Kits" for Teachers Participating in the Study

Boston College
Doctoral Dissertation
Procedural "Kits" for Collecting Teacher Assignments & Pupil Work

General details

Dear colleague

Thank you for agreeing to participate in my dissertation study.

I am exploring how Singapore Geography teachers seek to enhance pupil learning through their classroom assessments. To answer this question, I will be collecting samples of your classroom assessments, and your pupils' work, and conducting interviews with you from the period of April to August 2012.

A. Collection of Teacher Assessment and Pupil Work

I would like to collect three different sets of assessments which you use with your Secondary 1/2 Geography class.

First, please identify the following: (1) 6 students whom you consider to be in the high-ability group; and (2) 6 students whom you consider to be in the middle-ability group.

For each assessment, please help me collect 12 exemplars of your pupils' work. 6 of the assessments will be from the students you identified as high-ability, and the other 6 from students you identified as middle-ability. These assessments should be ones that have required your students to have completed individual written work.

Please note that you do not have to specially create assessments for this study. The details of the collection of teacher assessments and pupil work are on pages 1-2.

B. Interviews

I would like to meet you for three interviews, each expected to last between 50 min to 1 hour. I would like to schedule these interviews after you have marked the assessments that you collect for me.

To ensure that your thoughts about each assessment stays fresh in your mind, I would appreciate it if the interview could be held no later than 2 days after you finish marking and return the assessments. The interview questions are on pages 4-8.

If you have any questions, please email (karen.lam@bc.edu) or call me (98389431).

Thank you for your help and support.

Regards,

Karen Lam

Procedures for Collecting Teacher Assessments and Pupil Work

I. General

1. Please provide me with three geography assessments that you use with your Secondary 1/2 physical geography class. Ideally, the assessments will be ones that are implemented in April, May, and July 2012.

2. For each assessment that is submitted, you are requested to complete a cover sheet. The guide for this is presented below. Specifically, on the cover sheet, you will be asked to describe the nature of the geography assessment or task that you ask your pupils to do, the objectives, how the marks are assigned, your thoughts on the pupils' performance, and the type and nature of feedback that you will give them. Please be as detailed as possible because my analysis of the classroom assessments will depend on what you provide. Thank you.

II. Details for Completing Cover Sheet for Each of the Three Assessments

3. For each assessment, please complete the *Teacher Assessment Cover Sheet*. Each assessment has a different coloured cover sheet.
    a. Please provide details that you feel will help me understand the assessment and accompanying pupil work. Examples of such details include the explicit instructions given to pupils, marking scheme, rubrics, and lesson plan
    b. Attach the cover sheet with your completed details and the 12 pieces of pupil work. I will be providing you with one binder for each assessment.

III. Details for Collecting Teacher Assessments

4. Between April and August 2012, please collect three geography assessments that you administer to your Secondary 1/2 class. **They will be numbered Geography Task 1, Geography Task 2, Geography Task 3 and should be assigned to students in the period April to mid-August 2012. Please let me have a clean copy, one that does not have any annotation or writing. The geography assessments that I would like to collect from you for the period April to August 2012 should have the following criteria:**

   a. Require your pupils do individual written work. If several drafts of work are involved, please let me have a copy of the final piece.

   b. Require your students to demonstrate higher-order thinking in geography. This task may be a culminating assessment that requires your pupils to demonstrate the understanding of geography at a higher level, or integrates different aspects of one or more topics. Typically, this assessment may be one that is conducted after completing one or more topics or sub-topics. If the assessment has several steps, please let me have the final written piece.

Due to logistical and storage considerations, I am only able to study assessments that require students to complete written work individually.

IV. Details for Collecting Exemplars of Pupil Work

5. I will be collecting 12 exemplars of pupil work per assessment that you are sharing with me (from Section III). For example, for Geography Task 1, I would like to collect 12 pieces of pupil work. This is the same for Geography Task 2 and Task 3. Please make copies of the pupil work for me after you have completed marking. Ideally, your comments and/or marks should be on the marked assessments.

6. Please identify the following: (1) 6 students whom you consider to be in the high-ability group; and (2) 6 students whom you consider to be in the middle-ability group. For each task, select six high quality, and six middle quality pieces of student work.

7. To protect the identity of your pupils, I will provide you with Identification stickers to cover their names. Please take precautions not to cover any part of your pupils' responses, your comments, or your marks when applying the Identification stickers. The identification number assigned to each student should remain the same for each assessment submitted.

8. Please indicate High (for High-ability) or Medium (for Medium-ability) on each pupil's response as follows:

    a. High 1, High 2, … High 6 – for the six pieces of High quality work

    b. Medium 1, Medium 2, … Medium 6 – for the six pieces of Medium quality work

Details for Scheduling the Interviews

9. I will be meeting you for three interview sessions. Each interview will take place after you have marked and returned the assessments.

    a. Please call (98389431) or email (karen.lam@bc.edu) me once you have finished marking the assessments to schedule the interviews.

    b. Ideally, I would appreciate it if the interview could take place within two days after you finish marking the assessments. This is to ensure that your impressions about the pupil work remain fresh after marking.

Teacher Background Details

1. How old are you? Please check **one**.	☐ Under 25 ☐ 25-29	☐ 30-39 ☐ 40-49	☐ 50-59 ☐ > 60
2. By the end of this school year, how many years will you have taught (a) geography and (b) in total? Years teaching geography: _____ Total years in teaching: _____			
3. Gender	☐ Female	☐ Male	
4. Geography specialization:			
		Undergraduate: ☐ Yes ☐ No	
		Teacher training: ☐ Yes ☐ No	
5. What is the highest level of formal education you have completed? Other: _____		☐ Pre-university ☐ University (Bachelors) ☐ University (Honours)	☐ Masters ☐ Doctoral ☐ Teacher Education
6. Is there any background information about the class that I should know about when examining the assessments and pupil work?			
Note to interviewer: • Administer Informed Consent during Interview 1			

Geography Task 1 / 2 / 3 Cover Sheet

Please feel free to write or complete using a word processing programme.[35]

Class: Secondary 1/2 ☐ Express ☐ Special ☐ Normal (Academic)

Is this the end-of-topic assessment: ☐ Yes ☐ No

Design of assessment: ☐ Teacher's own ☐ Other sources (e.g., Internet, etc). ☐ Textbook _____

1. **Describe the assessment in detail and attach a <u>clean</u> copy of the assessment to this cover sheet.**

2. **What concepts, knowledge, or skills did you intend your pupils to demonstrate in this assessment?**

3. **Where does this assessment fit within the topic or theme in the syllabus?**

4. **How much time did it take for you to design (or if previously used or from commercial source, revise) this assessment?**

[35] This "kit" uses British spelling, as this is the convention adopted in Singapore.

5. How much time did you intend your pupils to spend on the assessment?

6. Did you inform your pupils that they may receive help of any kind when completing this piece of work?

☐ Peers ☐ Parents ☐ Another teacher

☐ Your feedback ☐ Other(s): _____

Please provide details:

_____

7. How was this assessment assessed? If you used a rubric, please attach it.

8. What criteria did you use to decide which assessments are *high* or *medium* quality?

9. Based on your experience, how did your selected 12 pupils perform on this assessment? If they did not perform to your expectations, do you have any idea why?

10. What were your pupils' comments about the assessment? (Could be the 12 or the other pupils in the class)

Thank you for your support.

Appendix 3: Dissertation Interview Protocol

Introduction for Interview 1

To begin, I want to thank you for agreeing to participate in my study.

As I mentioned when I contacted you, this interview is being carried out as part of a doctoral dissertation. As a result, anything that I learn from this interview will be used solely for the research. Furthermore, I will take all possible precautions to ensure that your identity is kept confidential, and that the information you provide will only be used for research purposes.

Please note that you are free to abstain from answering any questions which you are uncomfortable with, as well as to terminate the interview if you so wish. Finally, as this research study is for academic purposes, it is not intended to evaluate you, your classroom practices, or your colleagues and pupils. As such, please feel free to be as open and honest as possible.

The interview questions will focus on the following areas:

a. Introduction (for Interview 1)
   - Any questions that you have about the study
   - Informed consent

b. General discussion about "assessment" and assessment practices (for Interview 2)

c. Discussion of follow up decisions after classroom assessment

For Interviews 1, 2, and 3, we will also discuss the assessment, your thoughts after marking the task, and your analyses of your pupils' work.

The first interview is likely to last between 1 hour and 1 ¼ hours because of the introductory discussions. The second and third interviews are estimated to last between 45 min and 1 hour.

Interview 1: Introduction and conceptions of assessment

1. Tell me more about yourself.
   a. How long have you been teaching geography in this school? What are your roles and responsibilities?
   b. Why did you choose to teach geography? What is the purpose and value of geographical education?
   c. What do you enjoy about teaching geography?
   d. What are your learning goals for your geography students?
2. What was/is assessment like for you as (a) a student, and (b) a teacher?
3. Can you recall a memorable episode where you experienced 'assessment' in your daily life?
4. What does "assessment" mean to you when you teach?
   a. Here is a set of blank cards. Please write down the first **six** words that come to your mind when you hear the word 'assessment'.

b. Why do you think you picked these words? What do these words mean to you?

5. Why and how do you assess your pupils?
   a. Describe how you plan a typical geography assessment. Why do you use this approach?

Discussion of assessment task

Thank you for sharing your assessments with me. Let's look at Task 1/ Task 2 / Task 3*. (Circle the appropriate task).

Details of assessment

1. Who designed or created this assessment task? Have you used it previously? Is this a typical or a challenging assessment? [Definition of typical and challenging – as used in the 'Information Kit' to teachers.]
2. What learning did you want to elicit from your students? What did you want to find out about what your students know or are able to do? What is the value of this skill or knowledge?
3. Why did you select / design this assessment? What did you want your pupils to accomplish from this activity?
4. How did the class perform? Why do you think they performed so?
5. What were students' comments, if any, about the assessment? How and why will you use these comments?

Discussion about marking criteria

6. How did you mark the assessment? What criteria did you use? Are there other criteria that you use when marking pupil work?
7. What were common mistakes / strengths that you identified from the work?
8. How do you intend to use the marks / information from the assessment?

Discussion of pupil work (Write the letters, 'H' or 'M,' on the students' responses.) Please ensure that no pupil names are on the sheets which you pass to me.)

9. Let's look at your samples of the high and medium-level work.
   a. Why did you choose these examples?
   b. What criteria did you use to select "high" and "medium" level work?
   c. How do the samples compare with the rest of the class? Is the work of these 12 pupils representative of the standards in the class? Is the work of these 12 pupils work what you expected?
10. Now that you have completed marking the assessments, what do you think your pupils have learned or not learned? What is the value of this skill or knowledge?
    a. How do you decide whether your pupils accomplished what you wanted them to derive from this assessment/activity?

Discussion of formative assessment practices

11. Looking back at the assessment, what aspects worked well in terms of the
    a. objectives (your goals)
    b. mode (e.g., performance, written)
    c. format (e.g., MCQ, structured essay)?
12. What aspects would you want to change about (a) the assessment, and (b) the teaching of the lesson? Why?
13. Let's look at the feedback and comments you wrote on the assessment.
    a. Is this your typical comment? Do you give other comments than the grade? Why did you write or provide this comment?
    b. What decisions / strategies / approaches do you take to follow up after this?
    c. What do you usually say to your pupils as a class, and as individuals?

Overall reflection after discussing Task 1 / 2 / 3 (Circle the appropriate task)

We have now discussed the task you designed. Please think about what you just discussed about the task, and what you intended when you first assigned the task to your class.

14. How far did the assessment reflect the kind of learning objectives that you wanted to promote in your pupils?
15. How far were the assessments aligned with learning? How would you correct the misalignment (if any)?

Conclusion

16. Is there anything else you think I should know to understand your conceptions of assessment / assessment practice better? (can relate to the task, the pupil work, the interpretation and analysis)
17. Is there anything you would like to ask me?

Thank you for your candid, thoughtful, and deep sharing.

Interview 2: Classroom assessment practices

1. In 1997, Singapore launched Thinking Schools Learning Nation. It emphasizes higher-order thinking. Subsequently, Teach Less Learn More, launched in 2005 focuses on less content and deeper learning. This chart illustrates how TLLM envisions classroom assessment.
    a. What is your conception of HOTs in geography?
        i. What are some of the HOTs?
        ii. Can you describe what a geography assessment with HOTs would look like?
        iii. What is the role of higher-order thinking in your assessments?
        iv. What types of HOTs do you focus on?
2. These are some assessment practices that I found looking over the geography syllabus: multiple-choice questions, structured questions, oral presentations, portfolio, fieldwork, semestral assessment, continuous assessment.
    a. If I were to visit your school, what assessment types would I see being used (a) most and (b) least regularly in your geography classroom? Why?
    b. Are there assessment practices missing from this list that you use in your classroom? What are they? How are they used?
    c. What role do formal assessments play in the day-to-day basis in your teaching?
        i. How has this changed over time?
    d. What role do informal assessments play in the day-to-day basis in your teaching?
        i. How has this changed over time?
    e. If you had the autonomy to create and use higher-order assessments for geography,
        i. what would spur you?
        ii. what would hinder you?
3. Imagine you were at a tea session with the Minister, at town hall meetings, or at subject chapter sharing sessions. The topic of the day focuses on teachers' classroom assessment. What would be your reasons for creating and implementing the types of assessment that you typically use to assess your pupils?
        a. What are your reasons for these choices?
        b. If you could change classroom assessment in Singapore, what alternatives would you offer?
            i. Why do you prefer these options?
            ii. How can these options be realized?

Discussion of assessment task

Thank you for sharing your assessments with me. Let's look at Task 1/ Task 2 / Task 3*. (Circle the appropriate task).

Details of assessment

1. Who designed or created this assessment task? Have you used it previously? Is this a typical or a challenging assessment? [Definition of typical and challenging – as used in the 'Information Kit' to teachers.]
2. How does this assessment fit into a larger unit / syllabus / curriculum?
3. What learning did you want to elicit from your students? What did you want to find out about what your students know or are able to do? What is the value of this skill or knowledge?
4. Why did you select / design this assessment? What did you want your pupils to accomplish from this activity?
5. How did the class perform? Why do you think they performed so?
6. What were students' comments, if any, about the assessment? How and why will you use these comments?

Discussion about marking criteria

7. How did you mark the assessment? What criteria did you use? Are there other criteria that you use when marking pupil work?
8. What were common mistakes / strengths that you identified from the work?
9. How do you intend to use the marks / information from the assessment?

Discussion of pupil work (Write the letters, 'H' or 'M,' on the students' responses.) Please ensure that no pupil names are on the sheets which you pass to me.)

10. Let's look at your samples of the high and medium-level work.
    a. Why did you choose these examples?
    b. What criteria did you use to select "high" and "medium" level work?
    c. How do the samples compare with the rest of the class? Is the work of these 12 pupils representative of the standards in the class? Is the work of these 12 pupils work what you expected?
11. Now that you have completed marking the assessments, what do you think your pupils have learned or not learned? What is the value of this skill or knowledge?
    a. How do you decide whether your pupils accomplished what you wanted them to derive from this assessment/activity?
    b. How do you know if your pupils are learning? What counts as evidence for learning?

Discussion of formative assessment practices

12. Looking back at the assessment, what aspects worked well in terms of the
    a. objectives (your goals)
    b. mode (e.g., performance, written)

c. format (e.g., MCQ, structured essay)?

13. What aspects would you want to change about (a) the assessment, and (b) the teaching of the lesson? Why?
    a. What might you do differently for each of these students?
14. Let's look at the feedback and comments you wrote on the assessment.
    a. Is this your typical comment? Do you give other comments than the grade? Why did you write or provide this comment?
    b. What decisions / strategies / approaches do you take to follow up after this?
    c. What do you usually say to your pupils as a class, and as individuals?

Overall reflection after discussing Task 1 / 2 / 3 (Circle the appropriate task)

We have now discussed the task you designed. Please think about what you just discussed about the task, and what you intended when you first assigned the task to your class.

15. How far did the assessment reflect the kind of learning objectives that you wanted to promote in your pupils?

16. How far were the assessments aligned with learning? How would you correct the misalignment (if any)?

Conclusion

17. Is there anything else you think I should know to understand your view of ideal assessment? (can relate to the task, the pupil work, the interpretation and analysis)

18. Is there anything you would like to ask me?

Thank you for your candid, thoughtful, and deep sharing.

TLLM Vision	
More …	**Less …**
Remember why we teach	
For the learner	To rush through the syllabus
To excite passion	Out of fear of failure
For understanding	**To dispense information only**
For the test of life	**For a life of tests**
Reflect on what we teach	
The whole child	The subject
Values-centric	**Grades-centric**
Process	**Product**
Searching questions	**Textbook answers**
Reconsider how we teach	
Engaged learning	Drill and practice
Differentiated teaching	**'one-size-fits-all' instruction**
Guiding, facilitating, modeling	**Telling**
Formative and qualitative assessing	**Summative and quantitative testing**
Spirit of innovation and enterprise	**Set formulae, standard answers**

List of Assessment • Multiple-choice questions • Structured questions • Oral presentations • Portfolio • Fieldwork • Semestral assessment • Continuous assessment	Identify those most commonly used in your school (all subjects / geography) Identify those least commonly used in your school (all subjects / geography) What is missing from this list that you use?

Interview 3: Using, interpreting, and decision making with assessment data and reflections

Discussion of assessment task

Thank you for sharing your assessments with me. Let's look at Task 1/ Task 2 / Task 3*. (Circle the appropriate task).

Details of assessment

1. Who designed or created this assessment task? Have you used it previously? Is this a typical or a challenging assessment? [Definition of typical and challenging – as used in the 'Information Kit' to teachers.]
2. How does this assessment fit into a larger unit / syllabus / curriculum?
3. What learning did you want to elicit from your students? What did you want to find out about what your students know or are able to do? What is the value of this skill or knowledge?
4. Why did you select / design this assessment? What did you want your pupils to accomplish from this activity?
5. How did the class perform? Why do you think they performed so?
6. What were students' comments, if any, about the assessment? How and why will you use these comments?

Discussion about marking criteria

7. How did you mark the assessment? What criteria did you use? Are there other criteria that you use when marking pupil work?
8. What were common mistakes / strengths that you identified from the work?
9. How do you intend to use the marks / information from the assessment?

Discussion of pupil work (Write the letters, 'H' or 'M,' on the students' responses.) Please ensure that no pupil names are on the sheets which you pass to me.)

10. Let's look at your samples of the high and medium-level work.
    a. Why did you choose these examples?
    b. What criteria did you use to select "high" and "medium" level work?
    c. How do the samples compare with the rest of the class? Is the work of these 12 pupils representative of the standards in the class? Is the work of these 12 pupils work what you expected?
11. Now that you have completed marking the assessments, what do you think your pupils have learned or not learned? What is the value of this skill or knowledge?
    a. How do you decide whether your pupils accomplished what you wanted them to derive from this assessment/activity?

b. How do you know if your pupils are learning? What counts as evidence for learning?

Discussion of formative assessment practices

12. Looking back at the assessment, what aspects worked well in terms of the
    a. objectives (your goals)
    b. mode (e.g., performance, written)
    c. format (e.g., MCQ, structured essay)?
13. What aspects would you want to change about (a) the assessment, and (b) the teaching of the lesson? Why?
    a. What might you do differently for each of these students?
14. Let's look at the feedback and comments you wrote on the assessment.
    a. Is this your typical comment? Do you give other comments than the grade? Why did you write or provide this comment?
    b. What decisions / strategies / approaches do you take to follow up after this?
    c. What do you usually say to your pupils as a class, and as individuals?

Overall reflection after discussing Task 1 / 2 / 3 (Circle the appropriate task)

We have now discussed the task you designed. Please think about what you just discussed about the task, and what you intended when you first assigned the task to your class.

15. How far did the assessment reflect the kind of learning objectives that you wanted to promote in your pupils?

16. How far were the assessments aligned with learning? How would you correct the misalignment (if any)?

Theme Interview Questions

17. If you were to mentor beginning teachers to construct good quality assessments that assess higher-order thinking and learning,
    a. What would you tell them are key features that they should pay attention to?
    b. What should these young teachers know about how pupils learn that will help them to construct good assessments?
    c. What are ways for teachers to use and interpret assessment data? What strategies have worked for you? What decisions do you make instructionally after you complete marking?
18. Of the three assessments which you shared with me
    d. Which one most exemplifies the type of classroom that you most frequently use (e.g., in terms of format, objectives)
    e. Which one is the one you most want to improve on?
    f. Which one typifies the type of assessment envisaged in TSLN, TLLM?

g. Which one did your students find the most challenging?

Conclusion

19. Over the course of participating in this study,
    a. What new insights or reflections have you had about teaching and assessing geography? [To be asked at the end of Interview 3].
    b. How would you apply these new insights or reflections?
    c. In the course of participating in this study, what aspects of your assessment practices have you thought about or reflected on most? Why?

20. Is there anything else you think I should know to understand your view of ideal assessment? (can relate to the task, the pupil work, the interpretation and analysis)

21. Is there anything you would like to ask me?

Thank you for your candid, thoughtful, and deep sharing

Appendix 4: Invitation to Participate

Assessment in a knowledge society:
Preparing Singapore students for the "test of life," and not a "life of tests"
(A dissertation study exploring secondary Geography teachers' assessment practices)

Calling Geography colleagues

Have you asked yourself or wondered …

- What are my geography assessments like?
- Am I assessing my students well? How "good" are my assessments?
- How do I use assessment to enhance my students' learning, and develop critical thinking skills?
- How do I learn more about my assessment practices through reflection and discussion?

Are you interested in answering these questions? If YES, this study is right for you! ☺

Researcher: Wei Ling Karen Lam, Geography Teacher, Geography Teachers' Association member, PhD candidate at Boston College, USA

Changing conceptions of assessment in Singapore since 1997

Thinking Schools Learning Nation[36]

"develop creative thinking skills and learning skills for the future"
"encourage teachers and students to spend more time on projects that develop these skills"
Goh Chok Tong, TSLN Launch, 1997

Teach Less Learn More (2005)[37]

- ☞ More "**for understanding**" and less "dispensing information"
- ☞ More "**for the test of life**" and less "for a life of tests"
- ☞ More "**formative and qualitative assessing**" and less "summative and quantitative testing"

2013 Lower Secondary Geography syllabus[38]

- ☞ Continual assessments and Geographical Investigation

Your commitment

Point 1	Point 2	Point 3
April 2012	May 2012	July 2012
Teacher Interview 1 (1 hour)	Teacher Interview 2 (1 hour)	Teacher Interview 3 (1 hour)
Collection of artefacts	Collection of artefacts	Collection of artefacts

[36] Shaping our future: Thinking Schools, Learning Nation. Speech by Prime Minister Goh Chok Tong at the Opening of the 7th International Conference on Thinking on Monday, 2 June 1997, at the Suntec City Convention Centre Ballroom. Retrieved 1 February 2012 from http://www.moe.gov.sg/media/speeches/1997/020697.htm
[37] Teach Less, Learn More (BlueSky website). Retrieved 1 February 2012 from http://www3.moe.edu.sg/bluesky/tllm.htm
[38] Press article: http://www.channelnewsasia.com/stories/singaporelocalnews/view/1181941/1/.html

Interested in participating?

For further information, please e-✉:

Wei Ling Karen Lam (karen.lam@bc.edu)

Appendix 5: Authentic Intellectual Work Derived-Rubric (Teacher Assessments)

Background

1. You will receive 8 sets of assessments for Schools 001, 002, 003, 004, 005, 006, 007, and 008.
2. In each set, there are
    a. A blank copy of the assessment (Rate using "**Standards and Scoring Criteria for Teacher Assessment Tasks**)
    b. 12 sets of student work—6 from high ability students (e.g., for SCH001, these are numbered SCH001HA1, SCH001HA2, etc.), and 6 from middle ability students (e.g., for SCH001, these are numbered SCH001MA1, SCH001MA2, etc.)

       Please rate using "**Standards and Scoring Criteria for Pupil Work"**
    c. An answer sheet / marking scheme / rubric – Please use this in conjunction with (a) Standards and Scoring Criteria for Teacher Assessment Tasks.
3. For the standardization meeting on **Tuesday, 12 June 2012**, please use the following
    a. SCH001HA1, SCH002HA2, SCH003HA3, SCH006HA6
    b. SCH001MA5, SCH002MA4, SCH003MA3, SCH006MA2

Standards and Scoring Criteria for Teacher Assessment Tasks[a]

Key question: To what extent does successful completion of the task require the kind of cognitive work indicated by each standard?

The seven standards reflect three more general standards for authentic achievement as follows:

Construction of knowledge

- Organization of information
- Consideration of alternatives

Disciplined inquiry

- Disciplinary content
- Disciplinary process
- Elaborated written communication

Value beyond school

- Problem connected to the world beyond the classroom
- Audience beyond the school

General procedures (adapted from Koh, 2011)

1. If a task has different sections that imply different expectations (e.g., multiple-choice items, fill-in-the-blanks, short answer questions, and an extended question), the scores should reflect the teachers' apparent dominance or overall expectations. Overall expectations are indicated by the proportion of time or effort spent on different sections of the task and by criteria of evaluation if stated by the teacher.
2. Scores should take into account what students can reasonably be expected to do at their respective levels.
3. You should score only what you can see from the task. In determining the scores for each criterion, you should only consider the evidence in that specific written task.
4. When it is difficult to decide between two scores (e.g., 2 or 3), give the lower score. You will only give the higher score when a persuasive case can be made that the task meets the minimal criteria for the higher score.
5. Make your judgement based on the general intent of the criteria described in the introductory paragraphs.
6. The possible indicators given under each of the sub-criteria are by no means exhaustive, but rather merely a guide to assist your interpretation of the sub-criteria in the subject or content area.

[a] Extracted from Newmann, Secada & Wehlage (1995): A Guide to Authentic Instruction and Assessment: Vision, Standards and Scoring; RISER Manual (2001); and Koh, K.H. (2011). Improving Teachers' Assessment Literacy.

Standards and Scoring Criteria for Teacher Assessment Tasks

Standard and Descriptor[39]	Score
Standard 1: Organisation of information The task asks students to organize, synthesize, interpret, explain or evaluate complex information in addressing a concept, problem or issue. Consider the extent to which the task asks the student to organize, interpret, or evaluate complex information, rather than to retrieve or to reproduce isolated fragments of knowledge or to repeatedly apply previously learned algorithms and procedures. To score high, the task should call for interpretation of nuances of a topic that go deeper than surface exposure or familiarity. When students are asked to gather information for reports (that indicates some selectivity and organizing beyond mechanical copying), but are not asked for interpretation, evaluation, or synthesis, give a score of 2.	3 = high. The task's dominant expectation is for students to interpret, analyze, synthesize, or evaluate information, rather than merely to reproduce information. 2 = moderate There is some expectation for students to interpret, analyze, synthesize, or evaluate information, rather than merely to reproduce information. 1 = low There is very little or no expectation for students to interpret, analyze, synthesize, or evaluate information. The dominant expectation is that students will merely reproduce information gained by reading, listening, or observing. (*Adapted from RISER, 2001*)
Standard 2: Consideration of alternatives The task asks students to consider alternative solutions, strategies, perspectives, or points of view as they address a concept, problem, or issue. To what extent does success in the task require consideration of alternative solutions, strategies, perspectives and points of view?	3 = high 2 = moderate 1 = low

[39] Extracted from Newmann, Secada & Wehlage (1995). *A guide to authentic instruction and assessment: Vision, standards and scoring*; *RISER Manual* (2001); and Koh, K.H. (2011). *Improving teachers' assessment literacy*.

Standard and Descriptor[39]	Score
To score high, the task should clearly involve students in considering alternatives, either through explicit presentation of the alternatives or through an activity that cannot be successfully completed without examination of alternatives implicit in the work. It is not necessary that students' final conclusion include listing or weighing of alternatives, but this could be an impressive indicator that it was an expectation of the task.	
Standard 3: Disciplinary content The task asks students to show understanding and/or use of ideas, theories, or perspectives considered central to an academic or professional discipline. To what extent does the task promote students' understanding of and thinking about ideas, theories, or perspectives considered seminal or critical within an academic or professional discipline, or in interdisciplinary fields recognized in authoritative scholarship? Reference to isolated factual claims will not be considered indicators of significant disciplinary content unless the task requires students to apply powerful disciplinary ideas that organize and interpret the information.	3 = Success in the task clearly requires understanding of concepts, ideas, or theories central in a discipline. 2 = Success in the task seems to require understanding of concepts, ideas or theories central in a discipline, but the task does not make these very explicit. 1 = Success in the task can be achieved with a very superficial (or even without any) understanding of concepts, ideas, or theories central to any specific discipline.
Standard 4: Disciplinary Process The task asks students to use methods of inquiry, research, or communication characteristic of an academic or professional discipline. To what extent does the task lead students to use methods of inquiry, research, communication, and discourse characteristic of an academic or professional discipline? Some powerful processes of inquiry may not be linked uniquely to any specific discipline (e.g., interpreting graphs), but they will be valued if the task calls for their use in ways similar to important uses within the discipline.	3 = Success in the task requires the use of methods of inquiry or discourse important to the conduct of a discipline. 2 = Success in the task requires use of methods of inquiry or discourse not central to the conduct of a discipline. 1 = success in the task can be achieved without use of any specific methods of inquiry or discourse.
Standard 5: Elaborated Written Communication	4 = Analysis / Persuasion / Theory

Standard and Descriptor[39]	Score
The task tasks students to elaborate on their understanding, explanations, or conclusions through extended writing. This standard is intended to measure the extent to which a task requires students to elaborate on their ideas and conclusions through extended writing in a discipline. Expectations for elaborated communication can vary between the disciplines. <Adapted from those Newmann et al. (1996) created for Social studies>	The task requires explanations of generalizations, classifications and relationships relevant to a situation, problem, or theme. 3 = Report / Summary The task calls for an account of particular events or series of events, a generalized narrative, or a description of a recurrent pattern of events or steps in a procedure. 2 = Short-answer exercises Only one or two brief sentences per question are expected. 1 = multiple choice exercises; fill-in-the-blank exercises (answered with less than a sentence)
Standard 6: Problem Connected to the World Beyond the Classroom The task asks students to address a concept, problem, or issue that is similar to one that they have encountered, or are likely to encounter, in life beyond the classroom. To what extent does the task present students with a question, issue, or problem that they have actually encountered, or are likely to encounter, in their lives beyond school? Certain kinds of school knowledge may be considered valuable as cultural capital or cultural literacy needed in social, civic, or vocational situations beyond the classroom (e.g., knowing how a bill becomes a law). However, task demands for cultural valued, "basic" knowledge will not be counted here unless the task requires applying such knowledge to a specific problem likely to be encountered beyond the classroom. When students are allowed to choose topics of interest to them, this might also indicate likely application of knowledge beyond the instructional setting. But tasks	3 = The question, issue, or problem clearly resembles one that students have encountered, or are likely to encounter, in life beyond school. The resemblance is so clear that teacher explanation is not necessary for most students to grasp it. 2 = The question, issue, or problem bears some resemblance to real world experiences of the students, but the connections are not immediately apparent. The connections would be reasonably clear if explained by the teacher, but the task need not include such explanations to be rated 2. 1 = The problem has virtually no resemblance

Standard and Descriptor[39]	Score
that allow student choice do not necessarily connect to issues beyond the classroom. To score high on this standard, it must be clear that the question, issue, or problem which students confront resembles one that students have encountered, or are likely to encounter, in life beyond school.	to questions, issues, or problems that students have encountered, or are likely to encounter, beyond school. Even if the teacher tried to show the connections, it would be difficult to make a persuasive argument.
Standard 7: Audience Beyond the School The task asks students to communicate their knowledge, present a product or performance, or take some action for an audience beyond the teacher, classroom, and school building. Authenticity increases when students complete the task with the intention of communicating their knowledge to an audience beyond the teacher and when they actually communicate with that audience. Such communication can include informing others, trying to persuade others, performing, and taking other actions beyond the classroom. This refers not to the process of working on the task, but to the nature of the student's final product.	4 = Final product is presented to an audience beyond the school 3 = Final product is presented to an audience beyond the classroom, but within the school. 2 = Final product is presented to peers within the classroom. 1 = Final product is presented to the teacher.

Appendix 6: Interrater Agreement (Teacher Assessment)

Appendix 6
Interrater agreement (teacher assessment)

Measure	Exact agreement (%)	Exact or adjacent (%)
Standard 1	43.5	85.9
Standard 2	60.3	83.3
Standard 3	41.0	83.3
Standard 4	33.3	88.5
Standard 5	43.6	87.2
Standard 6	50.0	85.9

[a]There is no computation for Standard 7 because all of the assessments were created for the teacher alone, and thus received a score of 1 from each rater.

Appendix 7: Interrater Agreement (Student Work)

	Raters 1 and 2[a]		Raters 1 and 3		Raters 2[a] and 3	
Standard	Exact agreement (%)	Exact or adjacent (%)	Exact agreement (%)	Exact or adjacent (%)	Exact agreement (%)	Exact or adjacent (%)
1	46.7	84.3	46.5	89.9	65.6	100.0
2	56.3	90.6	46.2	92.7	62.5	96.9
3	59.3	100	68.1	99.3	56.3	100.0

[a] Note that R2 only scored 10% of the pupil work.

Appendix 8: Authentic Intellectual Work Derived-Rubric (Student Work)

General procedures (adapted from RISER, 2001)

The task is to estimate the extent to which the student's performance illustrates the kind of cognitive work indicated by each of the three standards:

- Analysis
- Disciplinary concepts
- Elaborated written communication

Each standard will be scored according to different rules, but the following apply to all three standards:

- Scores should be based only on evidence in the student's performance relevant to the criteria. Do not consider such as following directions, correct spelling, neatness, etc., unless they are relevant to the criteria.
- Scores may be limited by tasks which fail to call for social studies analysis, disciplinary conceptual understanding, or elaborated written communication, but the scores must be based upon the work shown.
- Take into account what students can reasonably be expected to do at the grade level. However, scores should still be assigned according to criteria in the standards, not relative to other papers that have been scored.
- When it is difficult to decide between two scores, give the higher score only when a persuasive case can be made that the paper meets minimal criteria for the higher score.
- If the specific wording of the criteria is not helpful in making judgements, base the score on the general intent or spirit of the standard described in the introductory paragraphs of the standard.
- Completion of the task is not necessary to score high.
- Scores may be limited by tasks which fail to call for students' authentic and intellectual performance, but the scores must be based only upon the work shown (Koh, 2011).
- Scores should take into account what students can reasonably be expected to do at their respective grade levels. In addition, scores should be assigned only according to "absolute" criteria, <u>not</u> relative to other pieces of student work that have been previously scored.

Standard and Descriptor[40] (Adapted from Newmann's social studies)	Score
Standard 1: Analysis Student performance demonstrates higher-order thinking with geography content by organizing, synthesizing, interpreting, evaluating, and hypothesizing to produce comparisons, contrasts, arguments, application of information to new contexts, and consideration of different ideas or points of view. This standard is intended to measure the extent to which students demonstrate cognitive activity that goes beyond mechanically recording, reporting, or otherwise reproducing information. Analysis may include proposing generalizations and supporting them with evidence; articulating and testing different theories or points of view; synthesizing and categorizing by applying abstractions to more specific information (this could include comparing similarities and differences); considering implications and application of information in new contexts; raising broad questions that help to interpret more specific information; or interpreting the meaning of personal roles, ideas, or events. The essential question is whether students demonstrate construction of knowledge through thinking and the organization of information, versus reproduction of knowledge by restating what has been previously given to them. The rhetorical form of students' statements might qualify as analysis (e.g., "the main reason for the American Revolution was taxation without representation"), but to score high on analysis, the student's work must appear to be reasonably original, not merely a restatement of some analysis that was given previously in a text or discussion. In assigning a 3 or 4, the rater should be (p.95) reasonably confident that no significant portion of the response has been virtually copied from some other source (i.e., text or oral statements of others). In scoring analysis, the proportion of work that illustrates analysis is more important than the actual number of statements indicating analysis.	4 = Substantial evidence of analysis. Most of the student's work includes analysis. At least three statements indicate that the student has successfully generalized, interpreted, tested, or synthesized specific information. 3 = Moderate evidence of analysis. A central portion of the student's work includes analysis. At least two statements indicate that the student has successfully generalized, interpreted, tested, or synthesized specific information. 2 = Some evidence of analysis. A small, but central, portion of the student's work includes analysis. At least one statement shows that the student has successfully generalized, interpreted, tested, or synthesized specific information. 1 = No evidence of analysis. Almost all statements consist of recording, or reporting specific information, without evidence of the student's organizing it or reflecting upon it; OR virtually all analysis offered is unsuccessful or in error.

[40] Extracted and adapted from Newmann, Secada & Wehlage (1995). *A guide to authentic instruction and assessment: Vision, standards and scoring*; *RISER Manual* (2001); and Koh, K.H. (2011). *Improving teachers' assessment literacy*.

Standard and Descriptor[40] (Adapted from Newmann's social studies)	Score
Standard 2: Disciplinary concepts Student performance demonstrates an understanding of ideas, concepts, theories, and principles from the social disciplines and civic life, and principles from the social disciplines and civic life by using them to interpret and explain specific, concrete information or events. This standard is intended to assess the extent to which students use important ideas of the social disciplines to make concrete information and events more meaningful. Substantive concepts (e.g., rock cycle, river erosion, migration) are often grounded in disciplinary knowledge from science and the social sciences, but many important ideas in geography (e.g., spatial distribution, time, change, patterns) may have no exclusive origins or associations within a single discipline. The main issue is the (p.96) extent to which the student has used substantive social ideas to organize, explain, interpret, summarize, and extend the meaning and significance of otherwise discrete pieces of information. Geography concepts may be used even though they may not be stated explicitly, and this might vary with student grade level. If the topic of the task is itself a substantive idea (e.g., revolution), students should get full credit for successful use of it. Give credit only for ideas that are used appropriately in the context of the assignment. No credit should be given for serious errors in application or interpretation. The phrase "geographical concepts" means a minimum of one geographical concept. The score of geography Disciplinary Concepts should be based on the quality of use of geography concepts, not on the proportion of student work that reflects geography concepts.	4 = The student has used geographical concepts to organize, explain, interpret, summarize, and extend the meaning and significance of otherwise discrete pieces of information. The use of ideas illustrates exemplary understanding. 3 = The student has included geographical concepts to organize, explain, interpret, summarize, and extend the meaning and significance of otherwise discrete pieces of information. The use of the ideas is somewhat limited and/or shows some flaws in understanding. 2 = Geographical concepts are included, but their use is significantly limited and/or shows significant flaws in understanding. 1 = The work includes virtually no geographical concepts, or the use of any that are included shows almost no understanding.

Standard and Descriptor[40] (Adapted from Newmann's social studies)	Score
Standard 3: Elaborated Written Communication Student performance demonstrates an elaborated account that is clear, coherent, and provides richness in details, qualifications and argument. The standard could be met by elaborated consideration of alternative points of view. To use the criteria, the scorer should identify specific points in the student work that are elaborated, and should make a judgment about the coherence of the overall framework in which various points are communicated. When a task includes several parts, the score for elaboration should be based on the part(s) answered in prose.	4 = Exceptional. The writer provides substantial and accurate elaboration for two or more important statements. The details, qualifications, and nuances are expressed within an overall coherent framework intended for the reader, and relevant to the topic. The response is so rich as to be worthy of display as an outstanding example of writing in geography. 3 = Elaborated The writer provides some elaboration for two or three important statements OR provides substantial elaboration for one important statement. In either case, the details, qualifications, and nuances are expressed within a coherent overall framework intended for the reader, relevant to the topic, and without major inaccuracies. 2 = Minimal The writer provides reasonably accurate elaboration for at least one important statement. 1 = Unsatisfactory The writer provides virtually no information or provides only disjointed details. OR the writer provides discrete claims, broad generalizations, slogans, or conclusions, but none are elaborated.