Measuring Multidimensional Science Learning: Item Design, Scoring, and Psychometric Considerations

Author: Courtney Castle

Boston College
Lynch School of Education

Department of
Measurement, Evaluation, Statistics, and Assessment

MEASURING MULTIDIMENSIONAL SCIENCE LEARNING:

ITEM DESIGN, SCORING, AND PSYCHOMETRIC CONSIDERATIONS

Dissertation
by

COURTNEY CASTLE

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

MAY 2018

Abstract

Measuring Multidimensional Science Learning:

Item Design, Scoring, and Psychometric Considerations

Courtney Castle

Dr. Henry Braun, Chair

The Next Generation Science Standards propose a multidimensional model of science learning, comprised of Core Disciplinary Ideas, Science and Engineering Practices, and Crosscutting Concepts (NGSS Lead States, 2013). Accordingly, there is a need for student assessment aligned with the new standards. Creating assessments that validly and reliably measure multidimensional science ability is a challenge for the measurement community (Pellegrino, et al., 2014). Multidimensional assessment tasks may need to go beyond typical item designs of standalone multiple-choice and short-answer items. Furthermore, scoring and modeling of student performance should account for the multidimensionality of the construct.

This research contributes to knowledge about best practices for multidimensional science assessment by exploring three areas of interest: 1) item design, 2) scoring rubrics, and 3) measurement models. This study investigated multidimensional scaffolding and response format by comparing alternative item designs on an elementary assessment of matter. Item variations had a different number of item prompts and/or response formats. Observations about student cognition and performance were collected during cognitive interviews and a pilot test. Items were scored using a holistic rubric and a multidimensional rubric, and interrater agreement was examined. Assessment data was

scaled with multidimensional scores and holistic scores, using unidimensional and multidimensional Rasch models, and model-data fit was compared.

Results showed that scaffolding is associated with more thorough responses, especially among low ability students. Students tended to utilize different cognitive processes to respond to selected-response items and constructed-response items, and were more likely to respond to selected-response arguments.

Interrater agreement was highest when the structure of the item aligned with the structure of the scoring rubric. Holistic scores provided similar reliability and precision as multidimensional scores, but item and person fit was poorer. Multidimensional subscales had lower reliability, less precise student estimates than the unidimensional model, and interdimensional correlations were high. However, the multidimensional rubric and model provide nuanced information about student performance and better fit to the response data.

Recommendations about optimal combinations of scaffolding, rubric, and measurement models are made for teachers, policymakers, and researchers.

Acknowledgments

This dissertation was the labor of several years, and would not have been possible without the support of many.

First, thanks to my committee. Dr. Nathaniel Brown trusted me with the enormous responsibility of overseeing an assessment development process from start to finish, and never ceased pushing me to grow. Dr. Henry Braun kindly and thoughtfully reminded me to consider the bigger picture. Dr. Kate McNeill was always willing to provide support and direction. And Dr. Sue Doubler, whose thoughtful leadership set the stage for this project.

The support of my classmates has been unwavering and appreciated. Victoria Centurino has been my constant companion throughout these past 7 years, and is probably the only person outside of my committee to read every word of this dissertation. The ERME wives have provided me with much needed friendship, advice, and distraction from the struggles of doctoral study.

Thanks also to my family. My parents, Mary and Gary, sacrificed and set high expectations for me, which ultimately led me to pursue this degree. My siblings have all contributed to this project in some way. Steffen helped me with programming and organizing hundreds of data files. Bryan generously helped me with some of the crucial, but tedious aspects of the analysis. Rachel was my constant cheerleader, providing emotional support and reminding me of my strengths. My in-laws, Jan and Chris, were voices of support and providers of much-needed getaway opportunities.

This dissertation is dedicated to my husband, Kevin, who is proud of me undeservedly, and beyond measure.

Table of Contents

Tables

Figures

Chapter 1: Introduction

Background

The dichotomy of science content and practice. Science is a multidimensional discipline, variously characterized over the past century as having multiple subdomains of essential knowledge and skills. Comprehensive science education includes coverage of science content – the facts that we traditionally associate with scientific knowledge, like the number of bones in the body, the atomic weight of carbon, or Newton's laws of motion – and science practice – the skills and processes utilized in the pursuit and application of scientific knowledge, like experimentation, data analysis, and modeling.[1] However, the relationship between content and practice in science education in the United States has changed markedly over the years, from a strict content-practice dichotomy, to the use of inquiry[2] in support of conceptual understanding, to the integration of content and practice to explain scientific phenomena (Barrow, 2006; Committee on a Conceptual Framework for New K-12 Science Education Standards, 2011).

Since the early 1990's, several major science education documents have included sections outlining the importance of scientific practice as a critical piece of science education in conjunction with scientific knowledge (e.g., Rutherford & Ahlgren, 1989;

[1] The NRC Framework (Committee on a Conceptual Framework for New K-12 Science Education Standards, 2011) introduced an additional subdomain called "crosscutting concepts" – the core concepts that are relevant across multiple scientific disciplines. Prior to this, discussions of multidimensional science mainly referred to content and practice, so the discussion here focuses on these two dimensions.

[2] Note that the terms "scientific inquiry" and "science practice" will be used somewhat interchangeably throughout this dissertation. These terms are slightly different; "inquiry" is used more often to emphasize the process of understanding a scientific phenomenon through observation and inference (National Science Teachers Association, 2004), whereas "practice" emphasizes the particular knowledge and skills that are utilized by scientists as they engage in the inquiry process (NGSS Lead States, 2013). However, there is considerable overlap between the two terms, and both terms describe the more general category of "doing" science, as compared to "knowing" scientific concepts.

AAAS, 1993; AAAS, 2001; National Research Council, 1996; National Research Council, 2000). However, the shift has been slow to trickle down to science education at the state, district, school, and classroom level. When inquiry is emphasized, it is often implicitly presented as a separate aspect of science with little connection to science content. Many state science frameworks include learning goals related to scientific practice or inquiry, but they are often separated from the content standards (e.g., Massachusetts Department of Education, 2006[3]; Florida Department of Education, 2015). This is similar to the approach taken in other academic disciplines: for example, the Common Core State Standards for Mathematics also include a separate section on mathematical practice, with only a few paragraphs to discuss the integration of practice with content (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010). Realization of an integrated relationship between science content and science practices thus remains a challenge for science educators.

Compounding the problem, science assessments at the state and national level largely emphasize and report science content knowledge only. Traditional science assessments have been designed to facilitate measurement of content; the items are designed to facilitate a focus on overall science proficiency, which is defined as science content knowledge. This focus is reflected in subsequent reports of student performance, which largely describe student proficiency with respect to science content. The result is a large amount of inertia as assessment developers must refresh their arsenal of science assessment strategies to accommodate the growing emphasis on knowledge in use (Pellegrino, Wilson, Koenig, & Beatty, 2014). Assessment is largely considered to be a

[3] Note that Massachusetts recently updated their science framework to reflect a stronger emphasis on the integration of science practices in science instruction (Massachusetts Department of Education, 2016).

motivational agent for instruction (Linn & Herman, 1997), and thus the failure of science assessment to account for knowledge and skills beyond just content exacerbates the division between science content knowledge and other dimensions of science learning.

The separation of content from practice has some undesirable side effects. Usually, this separation leads to an instructional emphasis on science content only, at the expense of science practices. Accordingly, low levels of emphasis on science inquiry have been reported among science textbooks (Chiappetta & Fillman, 2007; Eltinge & Roberts, 1993; Lumpe & Beck, 1996) and in science classrooms (Weiss, Pasley, Smith, Banilower, & Heck, 2003). When science practices are taught, they are often disconnected from content learning. Textbooks and science curricula frequently present science practices as a separate strand of learning goals, and there is little discussion of practices in the context of content (Lumpe & Beck, 1996). Even at the university level, science coursework is frequently separated into lecture courses and laboratory courses with disparate learning objectives (Kirschner & Meester, 1988). This results in a fragmented curricular structure in which students both learn about science content and use science practices, but rarely use science practices concurrently with learning about content.

The separation of content from practice misrepresents the discipline of science - in the real world, scientists and engineers simultaneously draw on science and engineering practices in the context of their conceptual expertise. Science as a discipline is not a static body of concepts and procedures, but a dynamic process in pursuit of knowledge about the world we live in (Carnegie Corporation of New York, 2009). Knowledge and practice contribute to each other - knowledge is uncovered through

practice, and knowledge contributes to practice. By keeping knowledge and practice separate, students fail to understand science as a process of active inquiry, in which questions are asked, investigations are designed, evidence is analyzed, arguments are formed, and knowledge is created (Duschl, Schweingruber, & Shouse, 2007).

The Next Generation Science Standards. In response to growing concerns about a) the overemphasis on content knowledge, and b) the isolation of content and practice in science instruction, science standards in the 1990's took a turn towards integrating practice into science instruction. (Rutherford & Ahlgren, 1989; AAAS, 1993; AAAS, 2001; National Research Council, 1996; National Research Council, 2000). These efforts culminated with the NRC Framework for K-12 Science Education (Committee on a Conceptual Framework for New K-12 Science Education Standards, 2011) and the subsequent 2013 release of the Next Generation Science Standards (NGSS) (NGSS Lead States, 2013). The NGSS were a collaborative effort, with direct input from 40 writers in 26 states and indirect input from many more researchers, scientists, educators, policymakers, and citizens.

In the Framework and the subsequent NGSS, explanation is set forth as the goal of the scientific enterprise, and investigation as the means to achieve that goal. As young children learn about the world, they are naturally curious. They crave explanations, and are able to reason about their observations and own ideas. The Framework and the NGSS leverage's children's early reasoning abilities to build on their early intuition about the world, starting with their initial misconceptions and simplifications and guiding them towards more sophisticated explanations of natural phenomena. The process of explanation-building provides the canvas for the development of science inquiry skills,

and is in turn enhanced and supported by inquiry as children progress through the grades Learning is a process through which questions are asked and answers are sought by the students; this enables students to experience the crux of "doing" science (Committee on a Conceptual Framework for New K-12 Science Education Standards, 2011).

In line with this vision, the NGSS outline the central tenets of science as a set of 3 interrelated dimensions: core ideas, practices, and crosscutting concepts. Disciplinary Core Ideas (DCIs) are the science concepts considered most essential for student learning. The Science and Engineering Practices (SEPs) are a list of 8 skills employed by scientists and engineers as they conduct research and design solutions to problems. Finally, the Crosscutting Concepts are 7 common themes across science - the big ideas that are applicable to multiple science topics and subtopics. All three dimensions are presented as a set of learning progressions, or descriptions of students' developing understanding over time, and the knowledge/skills present in one dimension support those in the other two dimensions.

A major emphasis of the standards is the integration of all 3 dimensions in instruction and assessment, thus facilitating "knowledge in use." According to the standards, "Scientific and Engineering Practices and Crosscutting Concepts are designed to be taught in context - not in a vacuum" (NGSS Executive Summary, Page 1). The NGSS aim to rectify the widespread disconnect of content and practice by explicitly encouraging their integration in a way that a) mimics how science is practiced in the real world, and b) draws connections between related concepts across science domains. Science learning is thereby an active process of drawing on known concepts and reasoning to generate scientific explanations in new contexts. The primary vehicle of the

NGSS is a set of performance expectations for student learning and assessment. The performance expectations describe the way that students should be able to apply knowledge and reasoning in new contexts by linking together core ideas, practices, and crosscutting concepts. The simultaneous incorporation of all 3 science learning dimensions in performance expectations is a clear signal that the pursuit of scientific explanations requires knowledge and practice to be utilized in conjunction.

Structure of the NGSS. The NGSS and the preceding NRC framework, on which the Standards are based, are organized into three dimensions: Disciplinary Core Ideas, Science and Engineering Practices, and Crosscutting Concepts (NGSS Lead States, 2013). These three dimensions capture a wide range of established scientific principles, ways of investigating and knowing, and transferable knowledge across scientific disciplines.

Disciplinary Core Ideas are the topics traditionally covered by science education, organized as a set of several learning progressions, which are sequences of progressively more sophisticated understandings of science concepts across grade bands. The learning progressions are organized into 3 topic areas: Earth and Space Science, Life Science, and Physical Science, and each of these topic areas contains several learning progressions related to specific science concepts. The number of scientific concepts represented in the learning progressions is purportedly less than in preceding documents, such that the resulting standards prioritize clarity and depth of understanding over breadth of content (NGSS Lead States, 2013).

The Science and Engineering Practices are essential skills related to science inquiry and engineering design. The practices are included in the NGSS because:

> Engaging in the practices of science helps students understand how scientific knowledge develops; such direct involvement gives them an appreciation of the wide range of approaches that are used to investigate, model, and explain the world. Engaging in the practices of engineering likewise helps students understand the word of engineers, as well as the links between engineering and science…The actual doing of science or engineering can also pique students' curiosity, capture their interest, and motivate their continued study; the insights thus gained help them recognize that the world of scientists and engineers is a creative endeavor – one that has deeply affected the world they live in. (NRC, 2012, p. 42)

Eight essential science and engineering practices are listed: 1) asking questions (for science) and defining problems (for engineering), 2) developing and using models, 3) planning and carrying out investigations, 4) analyzing and interpreting data, 5) using mathematics and computational thinking, 6) constructing explanations (for science) and designing solutions (for engineering), 7) engaging in argument from evidence, and 8) obtaining, evaluating, and communicating information (NGSS Lead States, 2013). Each of the practices is also presented as a progression, and students are expected to demonstrate mastery of successively more sophisticated scientific inquiry and engineering design skills at each grade band.

Finally, the Crosscutting Concepts present central scientific concepts that reappear frequently across scientific domains. There are 7 such critical concepts, which

"unify the study of science and engineering through their common application across fields" (NGSS, 2013, Appendix G, p. 1). These are 1) patterns, 2) cause and effect: mechanism and explanation, 3) scale, proportion, and quantity, 4) systems and system models, 5) energy and matter: flows, cycles, and conservation, 6) structure and function, and 7) stability and change. Again, the seven Crosscutting Concepts are laid out as progressions, such that student understanding of these thematic concepts is expected to increase in sophistication across time.

The three dimensions are then integrated to form multidimensional standards. Instead of content standards, which simply list the concepts and practices that each student should learn over the course of a given time period, the NGSS sets forth performance standards (the NGSS calls them performance expectations), which list in detail the types of behaviors that students must exhibit to demonstrate mastery of the standard.[4] Each performance expectation is related to a specific DCI, SEP, and CC, such that a certain level of proficiency on all three dimensions is required to demonstrate mastery of the performance expectation. See Figure 1.1 for an example performance expectation. This requirement, that at least one DCI, SEP, and CC be integrated into each performance expectation, is the most novel feature of the NGSS. By requiring the explicit integration of content, practice, and crosscutting concepts, the NGSS has emphatically prodded science education down a new path – a path in which knowledge, practice, and transferable concepts are intertwined.

[4] These types of standards are useful benchmarks for assessment, because they explicitly describe the evidence desired from a student's response to an assessment task.

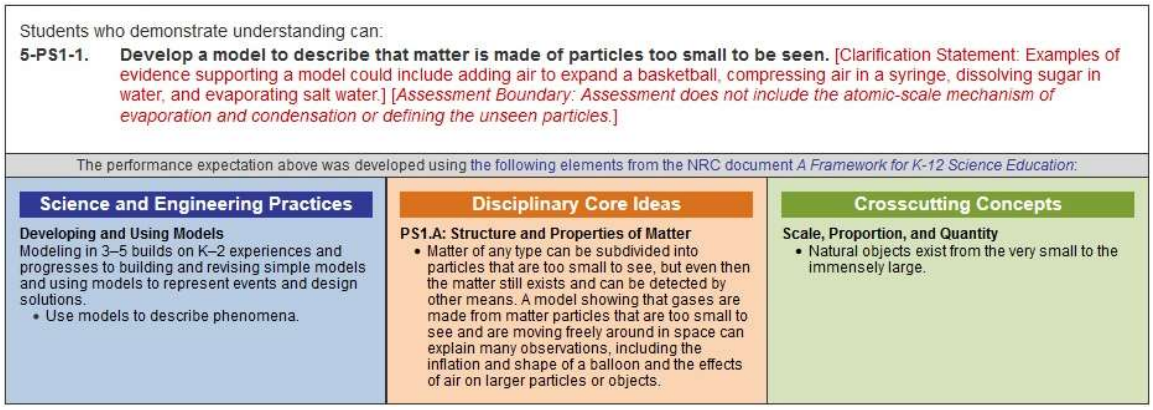


Figure 1.1. Example NGSS performance expectation. Retrieved from *www.nextgenscience.org.*

Statement of the problem

Accompanying the push for multidimensional science learning, as presented in the NRC Framework and NGSS, is a need for student assessment aligned with the new standards. Assessments are widely regarded as essential elements of educational reform, by:

> …[c]ommunicating the goals that school systems, schools, teachers, and students are expected to achieve; [p]roviding targets for teaching and learning; and [s]haping the performance of educators and students…assessments can motivate students to learn better, teachers to teach better, and schools to be more educationally effective. (Linn & Herman, 1997, p. iii)

Thus, the creation of assessments that reflect the new Next Generation Science Standards will be a critical component of the effort to integrate core ideas, practices, and crosscutting concepts in science classrooms.

The task of creating student assessments that validly and reliably measure student science ability along a progression of core ideas, practices, and crosscutting concepts is an unprecedented challenge for the measurement community. According to Pellegrino (2013), NGSS-aligned assessment should:

> ... help determine where a student can be placed along a sequence of progressively more "scientific" understandings of a given core idea that by definition includes successively more sophisticated applications of practices and crosscutting concepts. This is an unfamiliar idea in the realm of science assessments, which have more often been viewed as simply measuring whether students know about particular grade level content. (p. 320)

Traditionally, state assessments of science achievement utilize a unidimensional assessment framework that emphasizes content over practice. Furthermore, they primarily rely on multiple-choice and short-answer items, which offer the benefit of efficient and reliable measurement but fail to elicit some of the richer aspects of multidimensional science performance. For example, the Massachusetts science assessment blueprints neglect to mention scientific practice entirely (Massachusetts Department of Elementary and Secondary Education, 2015). Illinois, on the other hand, explicitly includes scientific practice but uses only multiple-choice assessment items to measure them (Illinois State Board of Education, 2013). Overall, current science accountability assessments are ill-suited to adequately capture evidence of the complex

cognitive processes elicited by the integration of core ideas with practices and crosscutting concepts.

To address the challenges facing the science assessment community, the National Research Council released a report entitled "Developing Assessments for the Next Generation Science Standards." The report describes in detail the challenges associated with assessing multidimensional science learning, and suggests some strategies for policymakers and test developers (Pellegrino, et al., 2014). First and foremost, they recommend that the assessment of science learning be implemented as a *system* of assessments, rather than just a single testing occasion. A systems approach differs from traditional assessment strategies in that it encourages the implementation of assessments that guide and support student learning in the classroom, in addition to large scale assessments that monitor students' proficiency with science concepts and skills along a learning progression. This would effect a drastic change in the way science learning is taught and assessed in the United States. The authors note:

> We see two primary challenges to taking advantage of this opportunity. One is to design assessment tasks so that they measure the NGSS performance expectations. The other is to determine strategies for assembling these tasks into assessments that can be administered in ways that produce scores that are valid, reliable, and fair and meet the particular technical measurement requirements necessary to support an intended monitoring purpose. (p. 138)

The first challenge refers to the design of multidimensional assessment tasks, in particular, "to design tasks that elicit the rich cognitive processes that define the hard-to-measure constructs as they were conceived and drafted by the standards' authors" (Gorin & Mislevy, 2013, p. 10). NGSS-aligned assessment items need to comprise complex, authentic tasks that allow students to demonstrate their ability to apply practices and draw connections between concepts in the context of grade-level appropriate science concepts.

In conjunction with (and facilitated by) a systems approach, Pellegrino, et al., recommend that new item types be considered.

> To adequately cover the three dimensions, assessment tasks will need to contain multiple components, such as a set of interrelated questions. It may be useful to focus on individual practices, core ideas, or crosscutting concepts in a specific component of an assessment task, but, together, the components need to support inferences about students' three-dimensional science learning as described in a given performance expectation. (p. 3)

Specifically, they describe performance-based questions (PBQs) as an example of the type of item that falls into this category. PBQs are extended prompts, in which students are asked to perform a series of related tasks that demonstrate their knowledge and skills (e.g., solve problems, create or use a model, design and interpret the results of an experiment). PBQs are rich, authentic tasks that truly fit the bill for integrating all three dimensions of science learning. However, there are drawbacks to developing assessments that rely heavily on performance-based questions, especially when the purpose of

assessment is to monitor student achievement across a large domain. PBQ's demonstrate a large degree of task-specific variance, meaning that a particular PBQ has unique features that may differentially affect student performance when compared to other PBQ's that assess the same topic. Task-specific variance makes test equating difficult. Without test equating, it becomes challenging to make comparisons of student performance across time – one of the major purposes of monitoring assessments (Pellegrino, et al., 2014). Relatedly, PBQ's demonstrate a large examinee-task interaction, such that the pattern of performance on individual tasks varies widely from person to person (Baxter, Shavelson, Herman, Brown, & Valadez, 1993). This makes it unlikely that a student's ability can be estimated reliably from a small number of tasks. Because PBQ's are complex, extended tasks within a particular content area, they require a large amount of testing time. Using several PBQ's to reliably monitor a broad range of topics further extends the amount of time required for individual student testing. Scoring must be done by human raters, which is time-consuming and costly. Considering all of these constraints, it will be prudent to supplement performance tasks with other, more traditional items, as another part of an assessment system.

Ideally, these supplemental items could include a series of smaller components, among which the NGSS dimensions are divided. The NRC report uses the term "multicomponent" to refer to the structure of test items that utilize multiple subtasks to sequentially assess core ideas, science and engineering practices, and/or crosscutting concepts (Pellegrino, et al., 2014). Even though the separate components of each item may refer to different dimensions and utilize different response formats, these individual puzzle pieces can be assembled to paint a coherent picture of student understanding

across the dimensions. Multicomponent tasks could feasibly include selected-response,[5] short-answer, or extended response components (Pellegrino, et al., 2014). Their discussion of multicomponent items in science assessment is largely hypothetical, leaving many questions regarding item assembly, specifically, how to cross response format with the NGSS dimensions and which content areas and skills can be assessed in combination using a multicomponent item structure.

The second challenge noted by Pellegrino, et al. refers to the need for psychometric approaches for scaling student proficiency that account for a) the use of complex multicomponent items, and b) three separate but integrated dimensions of science learning. Traditional psychometric models were designed to scale student responses on assessment items measuring unidimensional constructs, and thus are inadequate for the new generation of science assessments. There are two major priorities for deciding upon the most appropriate psychometric model: reporting needs and statistical considerations. Score reporting should reflect the goal of assessment (Pellegrino, et al., 2014), and in the case of the Next Generation Science Standards this means that a single score may not be sufficient to represent the 3-dimensional structure of the standards. Dimensionality is a central concern of NGSS assessment and reporting, so psychometric methods that account for multidimensionality should be explored.

This leads to the second psychometric consideration: statistical power. Multidimensional item response models have an additional set of considerations over

[5] From a design perspective, the use of some selected-response components may be desirable, as selected response items generally take less time for a student to respond. This allows for an increase in the total number of responses that can be observed from each student, which increases test reliability (Gorin & Mislevy, 2013).

unidimensional models: a larger number of parameters are estimated, thus requiring a larger sample size, more test items, and greater constraints on the distributions of persons and items (Gorin & Mislevy, 2013). Additionally, the type of complex, interrelated multi-part items described in the NRC report will likely require multiple items to be situated in the same context - a violation of conditional independence. There are IRT models to account for such a violation (testlet models, e.g., Wang & Wilson, 2005), but again, they are more complex than traditional IRT models. Ultimately, it will be prudent to determine task types and psychometric modeling techniques concurrently, as they mutually influence each other (Gorin & Mislevy, 2013).

The process of developing NGSS-aligned assessments is laden with many challenges - design challenges, psychometric challenges, and logistical challenges - each of which can be addressed when encountered separately, but concurrently they present a daunting assignment for the assessment community. The design of next-generation science assessments should not be rushed; rather it should be undertaken systematically and deliberately in order to meet the noble goal of guiding and supporting, instead of limiting student learning.

Purpose of the study

In response to the NRC Board on Testing and Assessment's call for development of science assessments that capture all three dimensions of the Next Generation Science Standards, support student learning, and provide valid and reliable information for monitoring (Pellegrino, Wilson, Koenig, & Beatty, 2014), this study contributes to the research evaluating different strategies for assessing multidimensional science proficiency. In order to assess science learning, we need to learn two things: 1) which

kinds of items are best-suited for assessing multidimensional science learning, and 2) what kind of measurement model is best-suited for describing multidimensional performance in science.

In order to capture student abilities on multiple dimensions of science learning, items must be carefully crafted to elicit responses that demonstrate the student's level of proficiency in three subdomains. It is unclear exactly how this should be done. In particular, should items use scaffolding to explicitly elicit separate responses for each dimension, and is it possible for selected-response items to clearly elicit evidence of student ability on multiple dimensions? This dissertation will explore each of these issues by comparing several indicators of assessment quality derived from alternative item designs that utilize a) varying amounts of multidimensional scaffolding, and b) open- and selected- response formats. It will also explore the impact of these design variations on students of varying ability levels.

To categorize and summarize student performance on individual items into a numerical estimate (or estimates) that describe overall performance, individual responses must be evaluated and scored and an appropriate measurement model must be selected. The chosen scoring rubric and measurement model will reflect the relationships among the assessment's Disciplinary Core Idea, Science and Engineering Practice, and Crosscutting Concept. This dissertation will also explore the relationships among the student abilities on the three NGSS dimensions of science learning by comparing different ways of scoring and scaling student responses. In particular, it will compare the appropriateness of using a) a unidimensional scoring rubric that evaluates overall item performance holistically along with a unidimensional measurement model, b) a

multidimensional scoring rubric that differentiates between the three assessment dimensions with a unidimensional measurement model, and c) a multidimensional scoring rubric with a multidimensional measurement model.

Finally, best practices dictate that any inferences drawn from student performance on any assessment should be supported by evidence that they are justified; this quality is called *validity*. In this case, the assessment is based on a theory about how students develop an understanding of science concepts related to matter, measurement, and argumentation. Student performance on the assessment is used to make inferences about their ability on the underlying multidimensional construct. Therefore, it is important to verify that the assessment tasks and the examinees' earned scores are indicators of the intended constructs, and that observations of student performance conform to the underlying theory about the progression of student ability (Kane, 2013). To verify that the assessment tasks adequately represent the intended construct, task development should involve input from experts who are familiar with the construct and relevant student behavior. Students' cognitive processes can be examined through interview to verify that they are utilizing the intended constructs during the tasks. Several psychometric criteria can be examined to explore the alignment between student performance on the assessment tasks and the underlying theory of the construct. The distribution of item difficulty estimates may be compared to the predicted difficulty based on the theory. Item fit indicates whether a particular assessment task demonstrates an abnormal pattern of responses, which may be the result of construct-irrelevant variance. Dimensionality analysis provides an empirical check on the strength of the observed relationship between dimensions, which can be compared to the theoretical relationship

laid out in the NGSS. Depending on the evaluation of validity, there may be a need to make changes to the way that the items are designed, the way that performance on individual items is scored and aggregated, or even in the way that the construct is defined. These changes, if necessary, are undergone in an effort to ensure that inferences reflect the underlying construct. This study will also examine several indicators of conformity between the assessment inferences and the three underlying constructs (Structure of Matter; Scale, Proportion, and Quantity; and Engaging in Argument from Evidence). The results of this evaluation will inform whether any changes to the assessment items, measurement model, and/or construct definitions are needed.

Multidimensional scaffolding. For the remainder of the dissertation, the term *multidimensional scaffolding* [6] will be used to refer to the support provided by the item to direct students' focus to each assessment dimension separately. Multidimensional scaffolding is one way to think about organizing multicomponent items (as described by Pellegrino, et al., 2014) in a way that simultaneously a) supports student performance on complex multidimensional tasks, and b) elicits clearer evidence of student ability relative to each dimension. The multidimensional scaffolding used in this study will take various forms, including an explicit reference to each dimension within the prompt, or the use of a separate item prompt for each dimension. These forms will be contrasted to items

[6] A discussion of the term "scaffolding" and its appropriateness for use in assessment can be found in Chapter 2.

without multidimensional scaffolding, which contain a single task prompt that integrates the three dimensions without explicit reference to them separately.

Research questions

Based on the general need for assessments that account for multiple dimensions of learning, the specific need for science assessments aligned with the Next Generation Science standards, and the persistent ambiguity about best practices for multidimensional assessment, the following research questions emerge. Research questions 1 and 2 inform item design by exploring alternative strategies for writing multidimensional items. Research question 3 informs the process of gathering and summarizing information from student responses through scoring and scaling. Research question 4 explores the assessment's construct validity.

Research Questions:

1. To what extent does multidimensional scaffolding affect the quality of information[7] gained from students' responses to multidimensional assessment items?
    a. Does the impact of scaffolding vary for students of different abilities?
2. In the assessment of students' argumentation ability, does the use of a selected-response item format affect the extent to which the enacted construct reflects the intended construct?
3. To what extent do unidimensional and multidimensional scoring and modelling approaches affect the empirical relationships among the assessment items, and

[7] Here, "quality of information" is a general term that encompasses several qualitative and quantitative indicators that student responses to an assessment item provide interpretable information that may be used to make robust inferences about their underlying ability. A complete list and rationale for the criterion used to evaluate "quality" for each Research Question may be found in Chapter 3.

what does this imply about the relationships between the 3 dimensions of science learning?

4. How well does student performance reflect the hypothesized definitions of the underlying constructs and their relationships?

Research context

The proposed research will take place in the context of the development of a 4th grade science assessment. The assessment is intended to collect summative information about student learning from the *Inquiry Project*: a science curriculum for Grades 3-5 that utilizes an inquiry-based approach to teach students about the nature of matter (TERC, 2011). The concepts of material, weight, and volume are the main focus of the curriculum. These concepts are iteratively explored throughout grades 3-5, ultimately forming a foundation for an understanding of the particulate model of matter in later grades. The curriculum was developed in conjunction with researchers and educators at TERC, an educational research organization in Cambridge, MA, and Tufts University. The curriculum incorporates core ideas, science practices, and crosscutting concepts from the NGSS, thus providing a suitable opportunity to employ multidimensional, NGSS-aligned items in assessment.

In line with the Next Generation Science Standards and the *Inquiry Project* curriculum, the assessment will focus on the learning progression PS1.A "Structure of Matter" as the primary Disciplinary Core Idea. In the 3rd grade *Inquiry Project* curriculum, students discuss the material composition of objects and are introduced to the concepts of weight and volume. In 4th grade, the curriculum specifically focuses on earth materials and introduces the concept of density with both solid and liquid materials.

Fourth grade students also discuss conservation of matter when an object is physically reshaped or broken into pieces. In 5th grade, students encounter the phase changes of water and consider whether air is matter. Students also observe conservation of matter when one substance is dissolved in another. The curriculum is structured so that it supports a beginning understanding of the particulate model of matter by the end of grade 5 (TERC, 2011). These curricular activities correspond well with the NGSS learning targets for PS1.A "Structure of Matter" for grades 3-5:

> Because matter exists as particles that are too small to see, matter is always conserved even if it seems to disappear. Measurements of a variety of observable properties can be used to identify particular materials. (NGSS Appendix E, p. 7)

The *Inquiry Project* curriculum contains elements of all 8 Science and Engineering Practices from the NGSS (TERC, n.d.). However, some of the SEP's play a more prominent role in the Inquiry Project curriculum than others. Therefore, rather than assessing all SEP's, this study will be limited to only one: Engaging in Argument from Evidence. Throughout the *Inquiry Project* curriculum, students conduct investigations and use the ensuing data to look at patterns in the weight and volume of objects. Using their observations as evidence, students learn about the relationship between material, weight, and volume through reasoning (TERC, 2011).

Again, the *Inquiry Project* incorporates elements of all seven Crosscutting Concepts listed in the NGSS (TERC). This project, however, will assess only one: Scale, Proportion, and Quantity. According to the NGSS, in grades 3-5 "Students recognize

natural objects and observable phenomena exist from the very small to the immensely large. They use standard units to measure and describe physical quantities such as weight, time, temperature, and volume." (NGSS Appendix G, p. 7) This Crosscutting Concept is a significant component of the Inquiry Project curriculum; as part of the prescribed activities, students learn how to measure the weight and volume of objects, represent relative weight and relative volume, explore the relationship between weight and volume, gain a rudimentary understanding of density, and use weight as evidence of conservation of matter. All of these activities require an understanding of size, weight, and volume, their respective measurements, and a sense of relative quantity.

The goals of the aforementioned *Inquiry Project* assessment are to a) obtain pre- and post- measures of student understanding for use in evaluating the impact of the curriculum and professional development, and b) provide items suitable for classroom assessment for teacher use at the end of the curricular unit. To accomplish both of these goals, items were inspired by the classroom activities contained in the *Inquiry Project* curriculum, and the concepts and practices therein. Items specifically target the benchmark levels of skill and understanding associated with the beginning and end of the 4^{th} grade curricular sequence.

Assessment development approach

The project's assessment development process is grounded in the *construct modeling* approach (Wilson, 2005; Kennedy & Wilson, 2007; Wilson, 2009; Brown & Wilson, 2011; Wilson, 2012), a model for assessment development based on four building blocks, or "cornerstones" (Brown & Wilson, 2011). These building blocks are (1) a clearly-defined construct, (2) items that require a demonstration of the construct, (3)

an outcome space, in which classes of responses are given scores according to how well they reflect proficiency in performing the task set forth by the item, and (4) a measurement model, which allows for summarization and interpretation of the information obtained from scored responses in relation to the underlying construct (Wilson, 2005). These four building blocks are often encountered sequentially in practice; however, the blocks are theoretically linked such that any particular stage in the process must be considered in terms of its impact on all remaining building blocks as well.

The process of assessment development involves an assumption that a student's response to an item is an indicator of her understanding of the underlying target construct, along with other relevant factors (e.g., general intelligence and literacy) and construct-irrelevant factors (e.g., test anxiety and test-taking strategies). The response(s) can then be used to make inferences about the degree to which the respondents understand the underlying construct. These inferences are affected by an underlying theory about how student understanding of the construct develops, and by the subsequent scoring and modeling decisions based on this theory (Wilson, 2005). This process is iterative, such that the results from several trials of the item(s) with many respondents provide empirical information about the nature of the underlying construct, in addition to information about the items and the respondents' understanding. This may lead to subsequent revisions of the construct theory, items, outcome space, and measurement model, creating a cycle in which an instrument is iteratively modified and refined as each of the four building blocks are brought into stronger alignment with one another (Brown

& Wilson, 2011; Wilson, 2005). Next, each of the four building blocks is discussed in turn.

Construct. A *construct* is difficult to define. According to Stenner, Smith, & Burdick (1983), constructs are "the means by which science orders observations." (p. 3). Anastasi (1986) called constructs "theoretical concepts of varying degrees of abstraction and generalizability which facilitate the understanding of empirical data." (p. 4-5). Kane (2006) defined constructs as "aspects or components of the postulated mechanisms or relationships" that account for observed phenomena (p. 42). In more basic terms, Wilson (2005) said that a construct is "an idea or a concept that is the theoretical object of our interest in the respondent." (p. 6). The common thread among all of these definitions is that constructs are intangible entities, but that our understanding of them must be facilitated by tangible observations. As assessment developers, our goal is to generate the observations that facilitate an evaluation of an individual's embodiment of the underlying construct. By examining the degree to which these observations conform to expected patterns, we can also contribute to general knowledge about the underlying construct.

To utilize a particular construct as the object of inference in assessment, it must be operationalized by putting forward some hypothesis about the relationship between the construct and one or more observable attributes (Kane, 2006). This hypothesis is inherent in the act of measurement: "The simple fact that numbers are assigned to observations in a systematic manner implies some hypothesis about what is being measured." (Stenner, Smith, & Burdick, 1983, pg. 3). However, an explicit statement of the construct theory helps to make clearer the ensuing inferences about respondents. Although this may seem intuitively obvious, the *definition* of the construct (i.e., specification of the construct

theory) is often taken for granted in the process of instrument development, with the brunt of development effort focused instead on creating, scoring, and scaling the items (Brown & Wilson, 2011).

The simplest operationalization of a construct is a unidimensional continuum with two extremes, such that the construct varies "from high to low, small to large, positive to negative, or strong to weak." (Wilson, 2005, p. 6). The definition of such a construct is captured in a *progress variable*, also referred to as a *construct map* (Wilson, 2005; Kennedy & Wilson, 2007; Wilson, 2009; Brown & Wilson, 2011; Wilson, 2012). A progress variable is a sequence of ordered qualitative descriptions, implying a continuum. Descriptive attributes hypothesized to correspond with less of the construct are captured at the bottom of the continuum, and those attributes hypothesized to correspond with more of the construct are captured at the top of the continuum. Levels of the construct are descriptive characteristics based on observations of some phenomenon in the real world. In the case of educational assessment, progress variables should be based on previous research about the construct and done in concert with domain experts, curriculum experts, teachers, and assessment developers (Wilson, 2005; 2012). That being said, the construct definition should not be considered as set in stone, as it will be reconsidered iteratively based on the empirical results of the assessment process (Brown & Wilson, 2011).

In educational measurement, the use of a progress variable to represent successive levels of mastery in relation to a construct implies that learning is a progression; that more sophisticated understanding develops over time, and that learning should be assessed in a way that allows for students to fall in the stages in-between a beginning and advanced understanding, instead of just comparing their work product(s) to a single

threshold of performance (Wilson, 2012). Progress variables are the building blocks of learning progressions, which are extended descriptions of the hypothetical development of concepts and skills over time within a particular domain of study (Wilson, 2009). The level of detail captured in a progress variable may vary, but generally it falls somewhere between the conciseness of state-level standards or frameworks and the detail of a particular curricular sequence (Wilson & Draney, 2004). Consequently, they function as convenient tools for developing assessments.

Some constructs are too complex to be operationalized in a single progress variable. For instance, some constructs have multiple dimensions or aspects, such as science learning as defined in the Next Generation Science Standards (NGSS Lead States, 2013). These constructs may be described by using a separate progress variable for each dimension (Wilson, 2005; Wilson, 2009; Wilson, 2012). This is the strategy utilized by this project, where the science knowledge supported by the Inquiry Project curriculum is divided among three progress variables for assessment: Structure and Properties of Matter; Scale, Proportion, and Quantity; and Engaging in Argument from Evidence.

Items design. In simplest terms, an item is an opportunity for a respondent to demonstrate their level of proficiency with respect to the construct of interest (Wilson, 2005). It may be as straightforward as a question asking for recall of facts, or as complex as a realistic situation that requires the coordination of multiple relevant and irrelevant pieces of information to perform a task. The fundamental characteristic of an item is that it elicits a response that contains evidence related to a student's mastery of an aspect of the construct. However, no single item will provide all of the necessary information about

a respondent (or about the construct). Furthermore, more items provide more information, which increases the precision with which respondents can be classified in relation to the to the construct. Therefore, multiple items are included on an instrument. The *items design* is a description of the set of items that is used to collect observations about a construct (Wilson, 2005).

Wilson (2005) describes the *construct component* and *descriptive components* of the items design. The construct component refers to the explicit link between each item and the construct map by specifying exactly which level(s) of the construct map are targeted by a particular item, as well as a justification explaining why the item will elicit a response that provides information about these level(s). This link is essential to the items design, as it provides the rationale for using each item to measure the construct. Descriptive components involve the non-construct related considerations of the items design, including whether the items will require self-report or performance, the response format, the specific context used to elicit the observation, the level of scaffolding in the item prompt, and many others. By necessity, some design components will be selected somewhat arbitrarily. The point of the items design, however, is to reduce the amount of arbitrariness as much as possible by making design specifications early in the development process.

The most oft talked-about design component is item format. Items may range from the completely open-ended – like observations, in which respondents may not even know that they are being observed – to the completely fixed – like multiple-choice or Likert-type questions. There are many types in between, including essays and oral or written short-response items. Wilson (2005) recommends that item development begin

with more open-ended formats before fixed-response variants are attempted, as this allows the designer to get a sense for the range of possible responses that appear when respondents are completely unrestricted. Then, decisions may be made about whether and how to constrain the responses.

Finally, the items' functionality should be investigated as part of the design process. This will help to both improve the items themselves and contribute to the validation of the resulting construct-based interpretations. The best way to examine the items' functionality is to examine the response processes utilized by the respondents as they encounter an item. This can be done via cognitive interview, a process in which an examiner asks the respondent to provide details about their thought processes either during their response or immediately after responding to an item (Wilson, 2005). The designer can then compare the respondent's reported mental processes to those processes specified in the construct component of the items design to verify the strength of the link between construct and item.

Outcome space. The outcome space describes the characteristics of a response that may be considered indicators of a particular level of mastery put forth in the construct theory. In practice, the outcome space is usually a rubric or scoring guide containing descriptions of the qualitative categories used to evaluate responses. It should consist of well-defined categories that clearly relate back to the theorized definition of the construct, with descriptions that are detailed enough to be meaningful and example responses where possible. These categories may be item-specific or may generally apply to all items measuring a construct under the given theory (Siegel, Nagle, & Barter, 2004). The number of categories should be large enough to account for all possible variations of

a response, but small enough that distinctions between categories reflect meaningful differences in responses (Wilson, 2005). One potential method for defining the categories of the outcome space is phenomenography (Masters & Wilson, 1997), in which responses are initially examined and qualitatively grouped, and category definitions and group membership criteria are iteratively revised by jointly examining the construct map and the responses. Another option is to use existing general categorization schemes (e.g., the SOLO taxonomy, Biggs & Collis, 1982) which may be adapted to suit a particular construct.

Scoring is the process of assigning values to each category of responses, such that higher values indicate that a response exhibits more of the construct and lower values indicate that a response exhibits less of a construct. Scoring may be pre-determined during item development, in the case of fixed-response items. However, if open-ended items are scored by raters, training procedures, including techniques like assessment moderation (Roberts, Wilson, & Draney, 1997), help to ensure that raters understand the scoring categories in terms of the construct and that scores are applied consistently across raters. The degree of convergence among raters (inter-rater reliability) is an indicator of how well the outcome space is defined.

Measurement model. The measurement model is the means by which scores are aggregated and related back to the hypothesized progression of the construct. The measurement model provides information about the respondent's overall placement on a particular construct, and may be used to guide decisions based on responses. Usually a mathematical or statistical model is used, with popular examples of measurement models

given by classical true score theory (CTT), factor analysis, item response theory models (IRT), and latent class models (Wilson, 2005).

In this study, IRT models, specifically, the Rasch family of item response models (Rasch, 1960; 1980) will be used as the measurement model. Item response theory models are probabilistic, meaning that rather than predicting the frequency of occurrence of a particular response category, the model predicts the *probability* of observing the response category. Item response theory models are useful because they provide information about both the items comprising a particular instrument and the respondents who responded to the items. In Rasch measurement, the Wright map is a visual representation of item and person information. Estimates of item difficulty and person ability are placed on a single common continuum, allowing for inferences about the construct based on information derived from their item responses. Thus, the measurement model brings the cycle of assessment full-circle, as a respondent's performance is summarized and interpreted in reference to the underlying construct (Brown & Wilson, 2011).

Prior *Inquiry Project* assessment development

Construct. Prior to the work described in this dissertation, progress variables were developed for three assessment dimensions: Structure of Matter; Scale, Proportion, and Quantity; and Engaging in Argument from Evidence. These progress variables were largely based on three NGSS learning progressions of the same name (NGSS Lead States, 2013).

The Structure and Properties of Matter progress variable (Figure 1.2) was largely informed by the learning progression underlying the Inquiry Project curriculum (Smith,

Wiser, Anderson, & Krajcik, 2006; Wiser, Smith, & Doubler, 2012), in addition to the NGSS progression. Although the Structure of Matter construct is presented as one continuous variable, it is decomposed into several different component concepts. These concepts evolve as the student's understanding strengthens. At the lowest level of the Structure of Matter progress variable, students' understanding of matter and material is limited to what they can easily perceive: size is "bigness" and weight is felt by heft. They can recognize certain materials by properties like color, smell, and taste. They are also able to recognize that the amount of material changes only when things are added or removed, but this does not necessarily correspond to weight or volume. At level 2, their understanding grows and they come to understand that tiny objects have weight[8] and take up space, even if it may be difficult to perceive. They start to link weight and amount of material, and they begin to recognize the different properties of solid, liquid, and granular materials. At level 3, they recognize the concept of volume as the amount of space take-up. As their understanding of weight and volume becomes more sophisticated, they begin to understand that different materials may be heavier or lighter than others, and develop a basic understanding of the relationship between the weight and volume of materials. At level 4, the relationship between weight and volume is formalized in the concept of density. They begin to grapple with ideas of weight and volume of imperceptible materials like gases. The idea of material properties begins to evolve and become a beginning understanding of the properties of chemical substances. Finally, these component concepts are finally subsumed into a larger understanding of the concept of

[8] Distinction between weight and mass does not develop until higher levels.

4	Matter	Matter has mass, weight and occupies space. Gaseous materials (air, water vapor) are recognized as matter. Solids, liquids, and gases are forms of matter; some materials can exist in all three forms (e.g., water). Beginning particulate model of matter: understands matter is made up of smaller particles of different substances.
	Substance	Substances are defined by properties (e.g., boiling points, melting points) and are invariant across phase change. Different substances composed of different particles.
	Mass	Mass is a measure of the amount of matter.
	Weight	Gases and other invisible pieces of matter have weight. Weight is proportional to mass (in a given gravitational field). Weight is invariant during phase change (freezing, melting, evaporating, condensation). Weight invariance can be used as evidence for conservation of matter.
	Volume	Distinguishes corpuscular volume from amount of space occupied by object. The latter is not conserved during phase change. Gases are much less dense than solids or liquids, and are compressible.
	Density	Now understood more formally as a relationship between mass and volume. The same material can have different densities in different states.
3	Material	Integrates weight, volume, heaviness for size in compositional model of materials. Knows any solid, liquid, or granular sample, however small, has weight and volume.
	Amount of material	More strongly links weight and volume with amount of material. Thinks tiny pieces weigh something and take up space because they are something.
	Weight	Knows tiny things have weight; weight is invariant across crushing and dissolving.
	Volume	Differentiates volume from area and understands volume of solid objects and liquids is invariant with reshaping. Understands water displacement depends upon volume of submerged object.
	Heaviness for size	Solid and liquid materials are (more or less) heavy for size. Differentiates heavy objects from heavy materials qualitatively.
2	Material	Understands that objects are constituted of materials, not just constructed from them, and can apply this compositional model to solid, granular and liquid materials. Different materials have different properties, and these properties make different materials suitable for different purposes. Recognizes that solids, powders and liquids have deep commonalities: they can be seen, touched, and felt, and are composed of little pieces of a given material.
	Amount of material	Has initial links between weight and amount of material in some contexts. Understands amount of material remains invariant with crushing and dissolving.
	Weight	Weight is more than just heft – it is an objective quantity that is related to amount of material. It is Invariant across reshaping (but not crushing).
	Size/ Occupied space	Understands even tiny things take up space.
1	Material	Knows the names of some liquids and solid materials and associates materials with some intensive properties (color, smell, taste) and properties associated with state (e.g., hardness, runniness). Knows solid objects are "made of materials" but thinks of "made of" as "constructed from" rather than "constituted of." Lacks compositional model.
	Amount of material	Has initial concept of amount of material as a quantity that remains invariant when object changes appearance with reshaping, because nothing is added/removed.
	Weight	Has perceptually based concept of weight centered on heft. Has crude generalizations linking weight and balance scale, and weight and size, but not amount of material. Weight varies when an object is physically rearranged.
	Size/ Occupied space	Has a general sense of how "big" an object is (judged perceptually), but no specific concept of volume as 3D measure of amount of space occupied by object. Knows that two solid objects cannot occupy the same space at the same time.

Figure 1.2. *Structure and Properties of Matter progress variable.*

matter as anything that has weight and occupies space. Matter is made of tiny pieces of different materials/substances.

The Scale, Proportion, and Quantity progress variable (Figure 1.3) was also inspired by the Inquiry Project curriculum (Smith, Wiser, Anderson, & Krajcik, 2006; Wiser, Smith, & of relative measures using addition, subtraction, and estimated proportional relationships. At the highest level of this progress variable, students have mastered methods of quantitative measurement for weight and volume, and can engage in more complex mathematical reasoning with these measurements (including precise proportional reasoning using multiplication and division).

The Engaging in Argument from Evidence progress variable (Figure 1.4) was informed by recent work on scientific explanation (Gotwals, Songer, & Bullard, 2012) and argumentation (Berland & McNeill, 2010; Osborne, et al., 2016). This progress variable is much simpler than the other two dimensions, and describes the student's developing ability to support statements with appropriate evidence and reasoning. At the lowest level of the construct, students are unable to articulate claims, much less support them. Beyond that, students may articulate claims but provide no evidence, or evidence that is not relevant to their claim. As students begin to support their claims, it is unlikely that they will jump to fully formed arguments with complete evidence and reasoning. Therefore, the middle level of the construct recognizes that students may successfully demonstrate *some* support for a claim, but fail to provide both components of a well-supported argument. After students learn to support arguments with relevant evidence and sound reasoning, they may also learn to use these components to refute opposing arguments – this is the highest level of the construct.

4	Use addition and subtraction to mathematically reason about volume (e.g., infer volume from change in water level, use water displacement to measure the volume of solid, liquid, and granular materials). Uses multiplication and division to mathematically reason using measurements (e.g., infer the weight and volume of proportionally smaller/lighter or larger/heavier objects made of the same material.) Quantifies and compares object characteristics based on proportional relationships (e.g., calculates "heaviness for size" (or density) as a ratio of an object's mass and volume using standard units).
3	Uses addition and subtraction to mathematically reason about weight (e.g., infer the weight of liquids in a container). Understands measuring instruments vary in precision (e.g., some scales more sensitive than others). Understands the structure of 3D arrays, and uses centimeter cubes to measure volume by comparison. Has generalized knowledge of fractions (there are an infinite number of fractions between any two integers); number and measure line is a dense, and quantities form a continuum Beginning understanding of measured characteristics that involve proportional relationships (i.e., uses relative scales to describe speed and density when distance/weight or time/volume is held constant.)
2	Knows how to measure weight with balance scale, length with rulers, and area with tiles Understands measures using scales, rulers are more reliable than senses. Understands that measures are more precise than senses. Can use measures to evaluate relative scale (for weight, length, area). Understands the structure of 2D arrays, and can use centimeter cubes to measure area. Has knowledge of a few special fractions (1/2, 1/4); number and measure line is spotty.
1	Believes the senses tell the truth. (Weight and size are assessed holistically and perceptually.) Places more reliance on senses than inferences and relations in judging quantities. Thinks of weight and size categorically (it's big, heavy or small, light) rather than as quantities; at best they use crude relative scales (e.g., bigger and smaller; heavier and lighter; hotter and colder; faster and slower) to describe objects (with few gradations). Only familiar with counting numbers (no fractions).

Figure 1.3. Scale, Proportion, and Quantity progress variable.

All three progress variables were iteratively examined by content experts and revised prior to beginning the item development work described in this dissertation

Items Design. Two specific characteristics of multidimensional items design (multidimensional scaffolding and response format) are a major topic of exploration for the research. However, assessment tasks and contexts were drawn from preceding work on student understanding of matter. Specifically, item development drew on two existing

5	Student makes a claim that answers a question or problem and supports that claim with BOTH evidence AND explanation linking the claim/evidence to scientific knowledge, and also rebuts alternative claims with justification (i.e., explaining why evidence/reasoning is incorrect)
4	Student makes a claim that answers a question or problem and supports that claim with BOTH evidence AND explanation linking the claim/evidence to scientific knowledge. May also rebut alternative claims without justification (i.e., doesn't explain why evidence/reasoning is incorrect or restates their own argument).
3	Student makes a claim that answers a question or problem and supports that claim with EITHER relevant evidence OR reasoning based on scientific knowledge. May also rebut alternative claims without justification (i.e., doesn't explain why evidence/reasoning is incorrect or restates their own argument).
2	Student makes a claim that answers a question or problem. May also provide irrelevant evidence (i.e., evidence that does not support their claim) or weak evidence (e.g., authority, personal experience).
1	Student doesn't make a claim, or makes a claim that is not relevant to the question or problem at hand.

Figure 1.4. Engaging in Argument from Evidence progress variable.

sources: item prototypes for assessing student understanding of matter (Smith, Wiser, Anderson, and Krajcik, 2006), and a sequence of hands-on assessment activities previously used in interviews with elementary students (Smith, 2009). Items covered an appropriate range of content topics and hypothesized difficulty, in line with the Inquiry Project curriculum and the assessment progress variables.

Significance of the study

The potential impact of this project is multi-faceted. The most immediate significance of the study relates to its role in developing a tool to evaluate student learning associated with the Inquiry Project curriculum. Although the Inquiry Project includes several suggestions for formative assessment embedded throughout the curriculum, it lacks a formal summative assessment (TERC, 2011). A summative

assessment for the Inquiry Project benefits classroom teachers who implement the curriculum by reducing the burden on teachers to generate their own evaluation of student understanding. When paired with a baseline measure, summative assessment also provides evaluative information about the impact of the curriculum on student understanding and contribute to measuring the effects of further intervention efforts associated with the curriculum. Evaluation of educational research provides information about the effectiveness of a particular method or intervention at accomplishing its stated goal, which ultimately supports the improvement of educational outcomes.

On a higher level, the work conducted for this dissertation contributes to the literature that addresses measurement practices for multidimensional constructs; in particular, three-dimensional science learning as described by the Next Generation Science Standards. Specifically, this work addresses the feasibility of multidimensional scaffolding and the selected-response item format for multidimensional science assessment through an examination of validity evidence based on student response processes. This evidence produces information that may also apply to the assessment of other core ideas, practices, and crosscutting concepts from the NGSS, or other non-NGSS multidimensional achievement constructs.

This study will also contribute to measurement literature related to modelling and reporting of student achievement on multidimensional assessments – specifically, NGSS-aligned science assessments. It explores whether multidimensional ability estimates based on the theoretically multidimensional NGSS structure are suitable for reporting science achievement, or whether a single estimate of overall science ability would provide an equally informative but more concise description of individual proficiency. Evaluation of

an assessment's dimensionality is a form of validity evidence based on the instrument's internal structure, and this type of validity evidence reflects back on the hypothesized progression of the construct (Wilson, 2005). In this case, the instrument's internal structure provides information about the extent to which the three-dimensional framework for science learning presented in the NRC Framework and the NGSS is empirically justified in this particular application.

Finally, student performance data provides some insight into the assessment's construct validity – the extent to which the constructs measured by the assessment correspond to the three constructs defined earlier in this chapter (Structure of Matter; Scale, Proportion, and Quantity; and Engaging in Argument from Evidence). Student performance provides evidence in support of or opposition to the construct definitions proposed in the three assessment progress variables, although the evidence may vary according to other factors in the students' educational background (e.g., group membership or instructional experiences). Thus, results inform the body of literature related to children's cognitive development of matter-related concepts, measurement, and argumentation.

Chapter 2: Literature Review

This literature review has two main parts. In the first, I provide general background on the topics addressed in the three research questions proposed for this dissertation: multidimensional scaffolding, response format, and scoring/reporting measurement data from multidimensional constructs. In all background sections, there is a specific focus on the topic as it relates to science assessment. Next, I describe several recent science assessments, with a particular focus on how test developers have approached the challenge of integrating content and practice in reference to the research questions.

Item scaffolding

Vocabulary. When the term *scaffolding* was first used by Wood, Bruner, and Ross (1976), it referred to a personal interaction between a one-on-one tutor and a child. They described the type of help provided by the tutor as engaging the child, directing the child to more manageable subtasks, keeping the child on task, and modeling the appropriate action. The term was later associated with Vygotsky's (1978) concept of the *zone of proximal development*, which describes the difference between a student's independent performance and their performance as aided by knowledgeable others. Overall, the initial discussion of scaffolding focuses on its role as an instructional tool in the service of student learning. One of the essential components of scaffolding is *fading*, (Pea, 2004) meaning that support gradually becomes unnecessary because the student has learned to complete the task unaided. Since the relationship between assessment and

student learning is less straightforward, it is unclear whether assessment supports align with the original conception of scaffolding.

Whether or not the proposed range of assessment supports[9] constitutes "scaffolding" as originally defined is up for debate. The proposed assessment supports share many similarities with the concept of scaffolding in its original formulation. Their main purpose is to provide structure for students as they complete a complex task. To achieve this, the item structure decomposes the task into smaller pieces, thereby reducing the degrees of freedom and focusing students to the most critical features of the task – all of which have been recognized as scaffolding strategies (Quintana, et al., 2004; Wood, et al., 1976). Furthermore, Quintana et al. (2004) note that students often have difficulties remembering to articulate their thought processes, and that scaffolding serves as a useful reminder to make their thinking transparent. In particular, scientific practices like explanation and argumentation require many pieces, and that scaffolding may help to guide students through this process (Quintana, et al., 2004). The proposed supports also serve this function, cuing students to make their reasoning explicit. On the surface, the described assessment supports seem to share many similarities with scaffolding.

In other regards, the use of scaffolding for assessment causes some divergence from the traditional conception of scaffolding as a support for learning. First, summative assessment is not instruction, and although the idealized role of summative assessment is to inform and improve instruction, any impact of summative assessment supports on student learning is several steps removed. Assessment supports do not only enable student performance, but also support complex inferences about three-dimensional ability

[9] See the section entitled "Item Design" in Chapter 3 for a detailed description of the proposed assessment supports.

based on student responses. This is an assessment-specific purpose, which does not align with the instructional focus of the original conception of scaffolding (Wood et al., 1976; Pea, 2004). Next, scaffolding is meant to be tailored to the students' abilities, supporting them where their understanding is lacking and fading or adapting as the student grows (Wood et al., 1976). On the other hand, standardization is often prioritized in assessment because it is an efficient way to directly compare student performance.

Pea (2004) argues that such supports are not scaffolding at all, but fall into a separate category called *distributed intelligence*, yet this term is not quite right either. Distributed intelligence describes tools that consistently enable some aspect of human performance by offloading some of the task burden onto an external artifact (e.g., calculators, computer software, and other tools that direct and simplify a task). However, assessment supports are not intended to enable higher levels of student performance, but only serve to enable clearer communication of the student's existing capability.

The application of the concept of scaffolding to assessment is tenuous at best, yet researchers and assessment developers have still adopted the term, using it to refer to item design components that enable task completion. As described in the following section, Songer and Gotwals (2012) applied the term to describe additional prompts that cue students to attend to relevant scientific concepts and components of scientific explanations, and the use of a selected response format to aid with explanation. They concluded that different levels of support are useful for gathering information about students of varying ability levels. By including items with varying levels of support within a single assessment, they were able to accomplish some degree of fading. Students who need more support would be able to accomplish the highly scaffolded items while

struggling with the less supportive items, and highly skilled students could still demonstrate their capability on less scaffolded items. The main difference between this type of scaffold-fading and instructional fading is that all assessment items are administered concurrently, rather than being gradually faded over time.[10]

The Science Assessment Item Collaborative (WestEd & CCSSO, 2015a; 2015b; 2015c) recently adopted the term *scaffolding*, as well, using it to describe the structure of prototypical multidimensional assessment items aligned with the NGSS. Scaffolding is described both as a feature that enables student performance on complex items and a communication tool:

> Scaffolding within and across assessment item clusters[11] helps guide students through a series of progressively more challenging interrelated questions, to better provide evidence of the knowledge and skills of students across a wide range of ability and understanding. (WestEd and CCSSO, 2015b, p. ii)

It is clear that the application of scaffolding in assessment, as demonstrated by both Songer and Gotwals (2012) and the Science Assessment Item Collaborative, diverges from the convention of scaffolding in instruction. Yet, both sources somehow demonstrate a shared understanding of scaffolding in assessment as a tool to enable

[10] Of course, assessment supports can be faded over time, but the current study design (a cross-sectional study of 4th and 5th graders) does not provide an appropriate context for exploration of this question.
[11] The term *item cluster* refers to a sequence of items that have been grouped together by a common context and scaffolding. The term is somewhat recognizable in assessment literature (e.g., Wainer & Kiely, 1987; Ferrara, Huynh, & Baghi, 1997), but discussion mainly focuses on psychometric concerns about conditional independence that arise when items share a common context. There is virtually no discussion about the impact of item clusters relative to single items, with regard to student response processes.

student attention and communication within complex assessment tasks. Therefore, it seems that Pea's (2004) criticism of usage of the term *scaffolding* outside of fading instructional supports falls on deaf ears among assessment developers.

Usage of scaffolding in formative and summative science assessment. Many researchers have explored the use of scaffolding in science learning and assessment as a tool to guide and strengthen student proficiency on a single dimension of student learning, but very few have evaluated the utility of scaffolding for multidimensional responses. Songer and Gotwals (2012) fall into this very small group based on research from their BioKIDS assessment (see pg 66 for more on BioKIDS). Building on previous work (Gotwals, 2006; Gotwals & Songer, 2010) they used four different types of items with various levels of scaffolding to assess students' ability to construct scientific explanations related to biology/ecology content. They used content scaffolds, explanation scaffolds, and a combination of both. They describe the "minimal" category of items as selected-response items containing both content and explanation scaffolds. These items provided students with evidence and asked students to select the claim best supported by the evidence. In open-ended items, students were provided with a scenario and asked to provide a scientific explanation for an observation. Content scaffolds took the form of hints that focused the students' attention on relevant features of the stimulus and/or reminded them of relevant science concepts. Explanation scaffolds split the task into three separate prompts that separately addressed the key components of a scientific explanation: claim, evidence, and reasoning. There were three variations of open-ended items: Intermediate I, which contained both content and explanation scaffolds;

Intermediate II, which contained explanation scaffolds only; and Complex, which contained no scaffolds.

They evaluated the effectiveness of each item type by examining the item information function, an IRT statistic that describes the inverse of the standard error for students of differing abilities. Items with smaller standard errors have larger information function values, meaning that they provide more precise information about student ability. They found that different levels of scaffolding were more effective at providing information, depending on the age of the students tested. At fourth and fifth grade, Complex items (no content or explanation scaffolding) were difficult for students, and did not provide as much information as Intermediate items (open-ended items with explanation scaffolding and content scaffolding). Minimal items (selected-response items with both content and explanation scaffolds) provided less information, at both grade levels. They hypothesize that this might be attributed to the amount of reading required by a highly scaffolded item. In sixth grade, however, the different scaffolding types fell into a pattern that corresponded with student ability, such that Minimal items provided more information about low-ability students, Intermediate items provided more information about average-ability students, Complex items provided more information about high-ability students.

In a later study (Gotwals & Songer, 2013) with an assessment focused on scientific explanation only, they again found that items with more scaffolding tended to be less difficult. To further examine this finding, they broke the explanation down into its component parts: claim, evidence, and reasoning. They found that the claim was the easiest part of an explanation, and scaffolded claims were not significantly more difficult

than non-scaffolded claims. Reasoning was the most difficult part of the explanation, and again, scaffolded reasoning was not significantly more difficult than non-scaffolded reasoning. However, evidence differed greatly in difficulty depending on whether the response was scaffolded or not. Scaffolded evidence was significantly easier than non-scaffolded evidence.

Based on the results of both studies, the authors suggested that scaffolding is most useful when students are faced with complex or unfamiliar tasks that lie just outside of their range of performance. In the case of multidimensional items, the integration of 3 dimensions of science learning into a single assessment or learning environment is a complex and unfamiliar requirement, especially for younger students. Therefore, these studies suggest that scaffolding may be necessary to guide students into explicitly attending to the three dimensions as part of their responses. The amount of scaffolding included in each item should be taken into consideration, as it may affect the item's ability to distinguish between students of high or low ability on one or more dimensions.

Elsewhere in science education research, researchers evaluated the use of scaffolds as tools to help students construct explanations or engage in argumentation, without explicitly considering the other content-related dimensions of science learning. Kang, Thompson, and Windschitl (2014) looked at high school teachers' use of scaffolding in science assessments. They identified five types of scaffolds: contextualized phenomena, rubrics, checklists, sentence frames, and drawings. After controlling for teacher and classroom characteristics using multilevel modeling, they found that three forms of scaffolding (contextualized phenomena, checklist, and rubric) were related to a higher quality (i.e., depth) of explanation from the students, but particularly

contextualized phenomena (i.e., making the elements of a scenario more familiar to the student). When multiple scaffolding techniques were used in conjunction, the scaffolding had a greater impact when contextualized phenomena was one of the techniques used. Chin and Teou (2009) used concept cartoons and scaffolding during formative assessment with 5th and 6th grade students in Singapore. They found that scaffolding students to make a claim about which character they agreed with and support their claim with reasoning required students to make their ideas explicit, where previously their ideas may not have been clearly articulated. This led to productive discussions between students and critical evaluation of other students' ideas. Again, these findings support the notion that scaffolding may support the production of high quality answers when students are presented with a complex or difficult task.

Item format

In simplest terms, items are generally described as having two types: selected-response and constructed-response (also known as open response or supply items) (Russell & Airasian, 2012). Selected response items do not require the examinee to provide any information of their own, but only to choose from one of the options given. The classic examples of this item type are multiple choice, true-false, and matching items. Open response items, on the other hand, require the examinee to provide additional information. The additional information may be as minimal as a word or number, or as large as an essay.

In recent years, researchers have begun to produce more elaborate taxonomies which account for nuance in the selected/constructed response dichotomy. Wilson (2005) lists a number of item characteristics, and the number that are predetermined before test

administration defines the item type and the level of constraint. For example, multiple-choice items require that all characteristics are predetermined before administration, including the response choices. Participant observations, on the other hand, do not require any preliminary consideration – not even necessarily the construct of interest. There are many gradations of fixation in between these two extremes. Scalise and Gifford (2006) list 7 different categories of items, from the *fully selected* (e.g., true/false and standard multiple choice) to the *fully constructed* (e.g., performance, portfolio). The five in-between categories are referred to as *intermediate constraint items*, and range from multiple choice with multiple answers to fault correction to short-answer and sentence completions to traditional essays. Obviously, the traditional dichotomy between selected-response and open-response items is not as straightforward as originally thought. Martinez (1999) notes that there is considerable variety among "constructed-response" items, with some items in this category requiring only simple recall, but others requiring more complex constructions.

The use of selected-response items in assessment, specifically multiple-choice items, is hotly debated. Critics of multiple-choice assessment question the alignment between the cognitive processes elicited by a multiple-choice item and the cognitive processes underlying the intended assessment domain (Gorin, 2006; Resnick & Resnick, 1996; Pellegrino, Wilson, Koenig, & Beatty, 2014). For example, Stanger-Hall (2012) found that anticipation of a multiple-choice item format was associated with lower-level cognitive engagement with study materials and a subsequent deficit in performance when a constructed-response item format was used in a university-level introductory biology course. This finding suggests that multiple-choice items elicit lower-level cognitive

engagement than their constructed-response counterparts and sheds some light on the unintended consequences of using a multiple-choice item format for assessment, especially when students have information about the test format beforehand. Many researchers counter-argue that multiple-choice items can assess more cognitively demanding skills if considerable effort is dedicated to the item writing process (Braun & Mislevy, 2005; Martinez, 1999; Pellegrino, Chudowsky, & Glaser, 2001). Nonetheless, it remains obvious that some skills fall outside the realm of multiple-choice, namely, those which require the generation of an original response (Martinez, 1999). This leads to skepticism that some aspects of science learning may be suitable for multiple-choice assessment, as scientific practice often involves generating arguments and justifications for claims, building models, and planning investigations – skills which inherently require creation. Furthermore, although multiple-choice items may viably assess complex skills, there is ostensibly a limit to the amount of complexity that multiple-choice items can handle. The use of multiple-choice items to assess multiple integrated dimensions may push past the limit of that viability.

Many researchers have explored this issue further, and have demonstrated shortcomings of multiple-choice items for assessing science practices. For example, Berg and Smith (1994) evaluated multiple-choice items about construction and analysis of graphical models by comparing results to near-identical constructed-response versions. They also conducted cognitive interviews with junior-high and high school students. They found disparities between the administration formats, such that students generally demonstrated higher levels of ability on constructed-response items than the supposedly easier multiple-choice items. Overall, they argue that multiple-choice items do not

provide a valid indicator of student understanding of graphs. Lee, Liu, & Linn (2011) compared multiple-choice items and "explanation items" which required students to explain or justify their response to a previous multiple-choice item. They found that the mean discrimination of the explanation items was higher than the multiple-choice items, and the explanation items captured a wider range of student ability than multiple-choice items. Additionally, explanation items were more sensitive to instructional intervention than multiple-choice items. They previously found (Liu, Lee, & Linn, 2011) that multiple-choice items were easier than constructed-response explanation items. All of these findings support an argument that constructed-response and multiple-choice items elicit different cognitive processes, and that the cognitive processes associated with constructed-response items are better suited for measurement of scientific practices.

Despite this, many researchers and educators have successfully used multiple-choice items in science assessment. The most successful examples of multiple-choice items innovate the format, conscientiously developing misconception-based distractors based on an underlying model of student understanding (Hestenes, Wells, & Swackhamer, 1992; Briggs, Alonzo, Schwab & Wilson, 2006; Liu, Lee, & Linn, 2011). The Force Concept Inventory is an early example of such an assessment. It assesses student understanding of Newtonian mechanics, and employs non-Newtonian misconceptions as distractors. The Inventory has been widely used with both high school and college students, and is generally regarded as a useful, reliable assessment of physics understanding (Hestenes, Wells, & Swackhamer, 1992). Briggs et al. (2006) took the idea of misconception-based distractors a step further by associating each distractor with a different level of student understanding. They found that these "ordered multiple-choice"

items were associated with higher levels of test reliability than traditional multiple-choice items. In both cases, deliberately crafted distractors lead to higher quality diagnostic information about student understanding. Liu, Lee, & Linn (2011) employed misconception-based distractors and used a two-tiered item format to create "explanation multiple choice" items. After responding to a multiple-choice item, students were asked to justify their initial choice by choosing an explanation from a list. Explanation multiple-choice items were easier than a constructed-response follow-up, which might indicate that they measure a different, slightly easier construct than explanation generation. However, the explanation multiple-choice format offers some advantages: it removes some of the ambiguity underlying student responses to traditional multiple-choice items, and provides an efficient alternative to constructed-response explanation items. In all, it seems that the use of multiple-choice items for assessment of science learning, particularly science practice, is a viable practice with many promising directions for improvement.

Scoring and reporting multidimensional constructs

A multidimensional construct is a latent domain that can be deconstructed into more than one subdomain. In theory, any construct can be endlessly divided into subcomponents until they no longer have any practical implication, so the real art is in being able to define multidimensional constructs that are empirically supported by assessment data and that are also substantively meaningful (Briggs & Wilson, 2003). Although a unidimensional definition of a construct is the default starting point for most measurement instruments (Wright & Masters, 1982), many educational assessment and

survey instruments have utilized a multidimensional framework to provide a more nuanced picture of student performance.

For example, the SAT assesses educational aptitude in three areas: Critical Reading, Writing, and Mathematics (The College Board, 2014c). Although performance on these three scales is often aggregated and reported as a single metric of performance, the overall score is built upon the three subdomains, each of which is measured by a set of content-specific items. Similarly, the ACT reports an overall measure of educational achievement derived from four subdomains: English, Mathematics, Reading, and Science (ACT, 2016). In each case, the separation between dimensions is obvious and extreme – both assessment programs refer to the sets of mutually exclusive items measuring the subdomains as "tests" in their own right.

Multidimensional assessment can also be found within a single educational domain (e.g., reading, mathematics, or science). For example, the 2015 Trends in International Mathematics and Science Study (TIMSS) reports four subdomains of Mathematics achievement at the 8th grade: Number, Algebra, Geometry, and Data and Chance. These dimensions are part of the larger construct of Mathematics, but they are mutually exclusive content areas that are measured by unique assessment items (Gronmo, Lindquist, Arora, & Mullis, 2013). The 2015 PARCC English Language Arts assessment reports scores for Reading and Writing, in addition to the overall ELA score. These subdomains are further subdivided to describe performance on Literary and Informational reading tasks (PARCC, 2015). Taken together with the examples in the preceding paragraph, these are only a few of the occasions where multidimensional constructs are used in educational assessment. These cases exemplify the default approach to the

assessment of multidimensional educational constructs. Here, the item pool is split among the dimensions, such that information about each dimension is measured by a unique set of items.

Most large-scale science assessments have a similar structure. Items are divided among science content domains (biology, physics, etc.) such that the dimensions are measured completely separately (e.g., TIMSS 2015, Jones, Wheeler, & Centurino, 2013). This approach is found among smaller scale instruments, as well – for example the NOSI-E measures five dimensions of the Nature of Science with five non-overlapping sets of items (Peoples, 2012). Altogether, this approach (Adams, Wilson, & Wang, 1997) rests upon the assumption that the measured dimensions are separable – an assumption that is called into question for the Next Generation Science Standards, where content, practice, and crosscutting concepts are explicitly described as *integrated* dimensions of science.

A common method for incorporating multiple assessment dimensions within items is what Briggs & Wilson (2003) describe as "a cross-referencing of sub-dimensions." For example, the 2009 NAEP framework describes standards for both content and science practices, and both dimensions are incorporated by crossing content statements with practice statements to create performance expectations, which are used as a starting point for item development (National Assessment Governing Board, 2007). Assessment items are then classified according to a science content domain – Physical Science, Life Science, or Earth and Space Science – and one of four science practices: Identifying Science Principles, Using Science Principles, Using Scientific Inquiry, or Using Technological Design (National Assessment Governing Board, 2010). Every item

can be mapped back to one content statement and one practice statement. Thus, items are divided among subdimensions of content and practice while also integrating the overarching dimensions of content and practice. This basically amounts to using all items on one scale for content and concurrently using all items again on another scale for practice. Several other large-scale science assessments also use this method to integrate multiple dimensions within items, including PISA 2015 (OECD, 2013).

Generally, assessments that utilize this strategy evaluate student responses by assigning a single overall science score to each response. This is the chosen scoring method for both the NAEP 2009/2011 and PISA 2012 assessments (National Assessment Governing Board, 2009; National Assessment Governing Board, 2011; OECD, 2012a). Thus, these assessments do not distinguish student ability on separate domains at the item level. However, reporting decisions about multiple domains differ from assessment to assessment. PISA 2006, which was the last PISA assessment focusing mainly on science, reported separate three subscores for each of two domains of science literacy: content and competency (OECD, 2008). In 2011, NAEP reported separate three subscores for science content only.

On large-scale assessments that utilize a concurrent between-item multidimensional approach such as NAEP and PISA, every item is included in the estimation of both a practice (or competency, or process, etc.) subscore and a content subscore, although the scoring process itself does not differentiate between the domains. The subscale scores therefore rest upon the assumption that a single item score can capture both content and practice (or competency, or process, etc.) without differentiating

between them. The role of the overarching dimensional structure, which divides science into content and practice (or competency, or process, etc.), is left unexplored.

Furthermore, science subscores from a concurrent between-item multidimensional approach are not always empirically justified by assessment data. For instance, Schwab (2007) attempted several multidimensional analyses of science data from PISA 2003. Along the lines of PISA's own analysis, she tried three different multidimensional models in accordance with the definition of science literacy in the PISA framework. In the first, items were divided into content dimensions (Physics, Biology, Earth Science); in the second they were divided among process dimensions (interpreting, understanding, and describing), and in the third they were divided among situation dimensions (earth and environment, life and health, and technology). Each time, the multidimensional model fit the data better than a unidimensional model in which all items were grouped together. However, the correlation between dimensions was so high (> 0.85) in each case that she ultimately concluded that a unidimensional model for overall science literacy was more appropriate.

Examples of assessments that measure science content and practice

Before the publication of the three-dimensional NRC Framework (Committee on a Conceptual Framework for New K-12 Science Education Standards, 2011) and Next Generation Science Standards (NGSS Lead States, 2013), many science learning documents and assessment frameworks began to emphasize the importance of science practices (or science inquiry[12]) in addition to science content (e.g., AAAS, 1993;

[12] As mentioned in Ch. 1, science *practice* refers to a set of skills (NGSS Lead States, 2013), whereas *inquiry* emphasizes ways of knowing (National Science Teachers Association, 2004). Here, inquiry and practices are used synonymously.

Massachusetts Department of Elementary and Secondary Education, 2006; National Assessment Governing Board, 2010). Since then, many science assessment programs have begun incorporating science practices as an essential component of assessments that traditionally only focused on content. However, definition of the relationship between content and practice, types of items, and the way that performance is reported differs from assessment to assessment. A few specific approaches are described below. These approaches vary widely in intended scope, assessment environment, and purpose (formative or summative). Altogether, they paint a picture in which both tradition and innovation have a large influence on science assessment practices.

The 2009 National Assessment of Educational Progress (NAEP).

Multidimensional scaffolding. The current NAEP framework (National Assessment Governing Board, 2007), which was the basis for every NAEP science assessment since 2009, includes both content and science practices. Assessment items are classified according to a science content domain – Physical Science, Life Science, or Earth and Space Science – and one of four science practices: Identifying Science Principles, Using Science Principles, Using Scientific Inquiry, or Using Technological Design (National Assessment Governing Board, 2010). Every paper-and-pencil item can be mapped back to one content statement and one practice statement, however, the item prompts themselves do not separate content and practice within the assessment task. Thus, the assessment development and data collection processes provide no relevant information about the utility of scaffolding for multidimensional assessment items.

Response format. The 2009 NAEP science assessment, for the most part, utilized basic item types that are typical of other similar large-scale paper-and-pencil assessment

programs: multiple-choice, and short or extended constructed-response items. On the main paper-and-pencil assessment, about 50% of student response time was devoted to multiple-choice items, with the remaining 50% devoted to constructed-response items (National Assessment Governing Board, 2007). The constructed-response format does have several variations (short and long response, predict/observe/explain, and concept mapping items), and they sometimes used item clusters that contained both multiple-choice and constructed-response items. However, for the most part these items were typical of traditional paper-and-pencil assessment formats. The major exceptions to this are hands-on performance tasks (HOTs), which provide students with lab equipment and materials and ask students to use them to solve a problem or answer a question, and interactive computer tasks (ICTs), which may ask students to search for information or conduct a simulated investigation in the context of a scientific problem (National Center for Education Statistics, 2012). These tasks were included because they are able to more clearly capture student ability to engage in scientific practice than short-answer and multiple-choice items (Fu, Raizen, & Shavelson, 2009). The hands-on performance tasks were administered to only a small subset of the students included in the national sample (National Assessment Governing Board, 2016; National Center for Education Statistics, 2011).

For the most part, NAEP relies on content experts to verify that assessment items measure the intended content and practice standards. In addition, a sample of NAEP items underwent classroom tryouts and cognitive labs prior to administration. Classroom tryouts allowed teachers and students to give feedback and make suggestions for item improvement. Cognitive labs investigated students' thought processes while completing

the items, and specifically focused on "the extent to which the science practices and cognitive processes (cognitive demands) evoked by the items are the ones intended by the performance expectation." (National Assessment Governing Board, 2007, p. 206) Note that they place a special emphasis on validating practices and cognitive demands, but not content. Furthermore, resource constraints allowed for only a sample of items to be subjected to cognitive labs, and priority was given to the innovative item types (hands-on performance tasks, interactive computer tasks, and concept maps) over constructed response and multiple-choice items. This means that relevant evidence about students' cognitive processes was rarely available for simple constructed-response and selected-response tasks, and when these were examined the focus was on only one of the two dimensions that the item purportedly assessed (practice).

Scoring and reporting. Items on the paper-and-pencil assessment were given a single score. Two-parameter, three-parameter, and generalized partial credit IRT models were used to estimate student ability on three content scales: Earth science, Physical science, and Life science (National Assessment Governing Board, 2009). Each of the content scales was assumed to be unidimensional and estimated separately. An estimate of overall science ability was calculated as a weighted average of the three content scores (National Assessment Governing Board, 2011). Although all items were classified according to both content and practice specifications and framework developers recommended reporting subscales for both content and practice (Fu, Raizen, & Shavelson, 2009), only content subscales were estimated and reported (National Center for Education Statistics, 2009). No reason was provided for this decision. HOTs and ICTs are not included in the scaling of the paper-and-pencil items due to unnamed technical

and practical reasons (Fu, Raizen, & Shavelson, 2009), but NAEP does provide special reports about HOT and ICT items (National Center for Education Statistics, 2012). Student performance on these items is summarized according to the percentage of students responding in each score category. Individual student performance on the HOT and ICT items was reported by taking the percentage of score points earned out of points attempted. Overall, the 2009 NAEP assessment results distinguish between content and practice qualitatively on occasion, and empirically not at all.

Programme for International Student Assessment (PISA) 2015.

Multidimensional scaffolding. In 2015, PISA measured a construct called "science literacy" which was defined by three areas: knowledge, competency, and system, each of which were subdivided into three subscales (resulting in 9 subscales total) (OECD, 2017). The competencies are three specific aspects of science practice: "explaining phenomena scientifically," "evaluate and design scientific enquiry," and "interpret data and evidence scientifically." Types of knowledge include content knowledge, procedural knowledge, and epistemic knowledge. The systems are the content domains of "Physical," "Living," and "Earth and Space." The competencies, types of scientific knowledge, and systems each provide three potential categorization schemes for defining scientific literacy as a three-dimensional construct subsumed by overarching categories of "competencies," "types of knowledge," and "systems," which themselves operate as higher-order dimensions of scientific literacy.

Assessment items were explicitly classified as belonging to one competency, one type of knowledge, and one system (OECD, 2017). Similar to NAEP (above), this results in a structure where the lower-order dimensions (i.e., the three categories of

competencies, types of knowledge, and systems) are assessed separately during assessment, while the higher-order dimensions ("competency," "type of knowledge," and "system") are integrated. Thus, content knowledge and procedural knowledge were never assessed together, for example, although content knowledge and designing scientific inquiry (a close relative of procedural knowledge) may be integrated in a single assessment task. Alignment between assessment items and the purported competency/type of knowledge assessed is verified through expert review (OECD, 2012a).

The item prompts themselves do not distinguish between the competency, type of knowledge, and system dimensions within the assessment task, and this conflation of dimensions may potentially obfuscate the intended constructs. Indeed, secondary analysis and review of previously administered PISA items has revealed discrepancies between the intended and actual constructs, notably, that the type of knowledge assessed is often hard to classify (Lau, 2009), that more than one type of knowledge may be assessed by a single item (Lau, 2009), that assessed and intended competencies may differ (Le Hebel, Montpied, & Tiberghien, 2014), and that some items purportedly assessing content may not actually require knowledge about content at all (Schwab, 2007).

Response format. The 2015 PISA science assessment was computer-based, and contained standard and interactive items (PISA, 2015). The standard items contained text passages, diagrams, and tables – stimulus materials that would easily be found on a paper-and-pencil exam – while interactive scenarios contained animations and simulations which allowed the student to actively *interact* with the stimulus. Regardless of whether the item stimulus was standard or interactive, student responses were

collected using the same formats: simple multiple-choice, complex multiple-choice (a series of yes/no items, multiple-response, fill-in-the-blank with options, and "drag and drop" ordering items), and constructed response. These three item types are roughly evenly distributed (OECD, 2012a), meaning that about 2/3 of the science items on PISA 2015 use some variation of a selected-response format. Although it is never explicitly stated, the heavy reliance on the selected-response format presumably serves to maximize the number of items on the assessment without increasing student testing time.

Scoring and reporting. On the PISA 2015 assessments, items were given a single score that ostensibly reflected a student's proficiency with the type of knowledge, competency, and system assessed. Scores on individual items were used to calibrate a unidimensional 2PL IRT model (OECD, 2017). Performance on the overall science scale is the major focus of the PISA science report, but PISA also reported data related to specific competencies, types of knowledge, and systems (OECD, 2017). These subscale scores were generated by recalibrating the model 3 times, each time examining the three subscales associated with only one of the higher-order dimensions (OECD, 2017). This had to be done because the IRT model only allows each item to be part of one subscale at a time. Since each item was assigned a classification on each of the three higher-order dimensions, the model had to be calibrated 3 separate times to generate plausible values on each subscale. A subscale average was computed, reflecting a country's relative performance on the subset of items classified as a "physical systems" item or an "explaining phenomena scientifically" item, for example.

Advanced Placement (AP) Biology, Chemistry, and Physics 2014-2015.

Multidimensional scaffolding. The AP science exam frameworks for chemistry, physics, and biology place a strong emphasis on the incorporation of inquiry and content (The College Board, 2014a; 2014b; 2015). The exam frameworks devote space to describing the content learning targets and 7 science practices separately, as well as presenting a set of learning objectives that incorporate both content and practice. The frameworks do not specify how the measurement of content and inquiry targets will be distributed across items, but all items are intended to demonstrate both content knowledge and practice – "Questions on the AP Biology Exam will require a combination of specific knowledge from the concept outline as well as its application through the science practices." (The College Board, 2015, p. 127). The combination of content and practice within an item is verified by expert review. However, content and practice are not distinguished within an assessment task.

Response format. The AP science exams are paper-and-pencil tests that utilize very traditional item formats – multiple-choice and constructed-response of varying lengths. For example, half of the current AP Biology exam is comprised of multiple-choice items and "grid-ins" (The College Board, 2015), which are essentially short constructed-response items that require a calculation. The multiple-choice items are rich, scenario-based items that require more than just memorization of content, and they may be linked together into item clusters. The remainder of the test uses long and short constructed-response items. The constructed-response items emphasize practice in the context of content, and often a long response item will require the use of more than one practice. The breakdown of item formats used in the AP Chemistry and AP Physics

exams is not identical, but very similar (The College Board, 2014a; 2014b). On the surface, these items are very high quality. The items explicitly incorporate science practices and content knowledge, are situated in realistic and engaging contexts, and ask students to utilize complex reasoning. However, the rationale behind the AP's design decisions is rather unclear. Despite the shift towards incorporating content and practice, no information is available about any effort to investigate the elicitation of content and practice-related cognitive processes associated with their items. Therefore, it is difficult to say for sure whether the College Board has succeeded at utilizing traditional paper-and-pencil item formats to assess complex, multidimensional constructs.

Scoring and reporting. Individual AP science items are given a single score that captures student proficiency with science content and practice simultaneously (The College Board, 2015). Neither multiple-choice nor open-response scoring criteria differentiate between the role of content knowledge and practice in an item response. Principles of Classical Test Theory are used to summarize student performance across items, and cut scores are developed to categorize students according to the overall level of proficiency demonstrated (Reshetar, 2012). AP exam scores are reported on a single scale from 1-5, with a score of 5 the highest level of proficiency and a score of 1 representing low proficiency (The College Board, 2014a, 2014b, 2015). This score provides information about both student content knowledge and student ability to engage in science practices, but relative ability on each dimension remains a mystery.

AP teachers are provided with a more detailed report, which provides information about multiple-choice performance across the content-related "Big Idea" subscales, calculated by summarizing information across items classified as measuring a particular

"Big Idea" (Reshetar, 2012). Teachers receive no information about their student's multiple-choice performance relative to the science practices. The teacher report also includes more detailed information about student performance on the individual criteria evaluated by the open-response scoring rubrics, but these criteria still do not provide information that distinguishes content knowledge from science practice.

Science Education for Public Understanding Program (SEPUP). The assessment system for the *Issues, Evidence, and You* middle school curriculum created by the Science Education for Public Understanding Program (SEPUP) was developed as part of the BEAR Assessment System (BAS) at UC Berkeley. The assessment system was based on a set of 5 interrelated variables: a science concept, two processes (designing and conducting investigations; evidence and tradeoffs), and two scientific skills (communicating scientific information; group interaction) (Roberts, Wilson, & Draney, 1997).

Multidimensional scaffolding. A single assessment item might measure a student's ability on one or more of the related variables targeted by the curriculum, but this was frequently done by using a single assessment prompt (Siegel, Nagle, & Barter, 2004). Therefore, a single student response was sometimes used to provide information about more than one dimension, without separating the dimensions within the task. The assessment also included multiple items linked together as part of a larger multi-component item, which essentially provides some degree of scaffolding (Siegel, Nagle, & Barter, 2004). The assessment developers did not address any differences that arose from assessing multiple dimensions with a single item compared to a multi-component item.

Response format. Several different item types are used, including multiple-choice (Briggs, Alonzo, Schwab, & Wilson, 2006) and open-response. Multi-component items utilized either or both response formats by pairing open-ended items, or by pairing a multiple-choice item with an open-ended item (Siegel, Nagle, & Barter, 2004). Multiple choice items were linked to only one dimension of the learning domain, but some constructed-response items provided information about three different dimensions simultaneously.

Scoring and reporting. Assessment items linked with multiple dimensions were given scores using a multidimensional scoring rubric, meaning that a single response produced multiple scores (Roberts, Wilson, & Draney, 1997). These scores were then used to estimate student abilities on each dimension using a multidimensional Rasch IRT model (Briggs & Wilson, 2003), which was shown to fit the data better than a unidimensional model. Multidimensional Rasch models provide information about student progression along the assessment dimensions in a way that provides distinct information about student ability on each dimension, but also acknowledges the relatedness of the assessment dimensions.

SimScientist. The SimScientist simulation (Quellmalz, Timms, Silberglitt, & Buckley, 2012) measures students' understanding of science content and practice in physical, life, and earth science domains. Six science practices are defined: use of science principles, prediction, observation, description, investigation, and explanation. Content and practice are considered separately as related dimensions, and each is defined by a progress variable with three levels of sophistication. The simulation contains several embedded assessments that provide formative information for teachers, as well as

"benchmark" assessments that are intended to provide summative information at critical transition points between units.

Multidimensional scaffolding. Each item and "observable event" (record of student activity in the simulation) was associated with at least one content or inquiry target. Because the items within a simulation all share the same context, each simulation acts as one large item with separate components related to content, practice, or both content and practice.

To ensure that items adequately elicited the intended content and practice targets, they underwent expert review, comparison to the 2009 NAEP framework and AAAS benchmarks, and cognitive interviews with students (Quellmalz, et al., 2012). During the cognitive interviews, they kept track of the number of times that students accessed relevant content and inquiry targets while completing a particular task, thus verifying that the tasks elicited the intended cognitive processes.

Results from the simulation-based assessment were compared with a paper-and-pencil assessment which used more traditional stand-alone items to measure both content and practice; however, their analysis focused on the empirical relationship between content and practice and did not explore how the difference in item structure might have affected performance. Therefore, the effect of the simulation's multi-component structure on student performance remains unclear.

Response format. SimScientist is a simulation-based assessment system, meaning that it is administered via computer and involves interactive components. Students can watch simulations and, in some cases, control simulations of real-life scenarios that allow the demonstration of scientific concepts. Assessment items include adapted paper-and-

pencil item types (e.g., multiple-choice, constructed-response), technology enhanced items (e.g., drawing arrows between components of a system), and informal "observable events" (tracked interactions with the simulation, e.g., variables manipulated, values assigned to variables, number of trials run). Student proficiency on content and inquiry targets was estimated using multiple methods at multiple points in time.

Scoring and reporting. Within the software itself, embedded assessment tasks were auto-scored according to a 3-point rubric which classified student performance as either "needs help," "progressing," or "on track." Teachers' reports included using the frequency of student appearance in each rubric category across all content and inquiry targets (Quellmalz & Silberglitt, 2012), allowing them to target instruction to address the specific concepts or science practices needed by a particular student. For the benchmark assessments, a Bayes Net was used to produce separate summative proficiency reports for both content and inquiry (Quellmalz, et al., 2012). Each student was classified on an overall 4-point proficiency scale for each content and inquiry category. To further examine the technical quality of the items, a between-item 2-dimensionsal (i.e., content and inquiry) IRT model was also fitted to student data from the benchmark assessment (Quellmalz, et al., 2012). Both the Bayes Net and IRT measurement models provide information about student performance on content and inquiry targets as separate dimensions, but both approaches also acknowledge and account for the relationship between student content knowledge and inquiry skills. In the IRT analysis, both dimensions had high reliability, and the discrimination between science content and practice was better (i.e., correlations between dimensions were lower) than a paper-and-pencil test of content knowledge and inquiry skills that was administered concurrently

(Quellmalz, et al., 2012). This suggests that a multidimensional analysis was appropriate for the simulation data.

BioKIDS.

Multidimensional scaffolding. The BioKIDS program at the University of Michigan has twin goals of improving middle school students' understanding of science concepts related to biodiversity and ecology, and teaching students how to construct scientific explanations. Biodiversity and scientific explanation are considered two separate, but related constructs, and each are described in a separate learning progression (Songer, Kelcey, & Gotwals, 2009; Gotwals & Songer, 2006a, Gotwals, 2006; Gotwals, Songer, & Bullard, 2012). At the conclusion of the curriculum, summative assessment tasks are administered. All items measure both constructs. Items were created with the aid of a content-reasoning matrix, in which progressively more complex science concepts are crossed with more sophisticated types of explanation (Gotwals, 2006; Gotwals & Songer, 2006a; Gotwals & Songer, 2006b; Gotwals & Songer, 2010). Each item inhabits a cell in the matrix, indicating the relative demand of content and practice for that item (e.g., low-content/high-reasoning or moderate-content/moderate-reasoning). Some items measured a third construct, analyzing and interpreting data, in place of constructing a scientific explanation. A similar content-reasoning matrix was used to design the data analysis items (Gotwals, 2006).

Student response processes were examined via think-aloud interviews with students, in which student responses were coded to reflect whether they referred to content knowledge, or used cognitive processes associated with constructing an explanation analyzing data (Gotwals, 2006; Gotwals & Songer, 2013). Nearly all items

elicited exactly the dimensions that were anticipated, and those that didn't were revised or removed.

As described in the previous review of scaffolding literature, items sometimes included content and explanation scaffolds, which provided clues about relevant content information and reminded students about the necessary components of a scientific explanation (e.g., Songer & Gotwals, 2012). Items with very high amounts of scaffolding (*minimal* items) and very low amounts of scaffolding (*complex* items) utilized only a single item prompt to gather assessment data on both dimensions. Items with an intermediate amount of scaffolding used separate prompts to remind students of the components of a scientific explanation. By comparing various levels of scaffolding, the authors found that the utility of the scaffolding varied depending on the ability level of the student, and that items with an intermediate level of scaffolding were generally the most informative across the board.

Response format. Assessment tasks were administered via paper-and-pencil, and asked students to select the appropriate response from a set of alternatives (multiple-choice), fill in the blank with a correct term or claim, or construct a scientific explanation based on a description of an event or a pattern of data. Selected-response and fill-in-the-blank items were found to be more informative for lower ability students, while open-ended items were generally more informative for students of average and high ability. This was true for both the scientific content and explanation dimensions, depending on the amount of scaffolding (Songer & Gotwals, 2012).

Scoring and reporting. The BioKIDS strategy for scoring items and subsequently summarizing student performance has varied over time. In some iterations of analysis,

items have received multiple scores for content, scientific explanation, and data analysis, and been fitted to a within-item two- or three-dimensional Rasch model (Gotwals, 2006; Gotwals & Songer, 2006a; Gotwals & Songer, 2010). Other times, the researchers prefer to score and report student performance on a unidimensional scale that merges content and explanation, citing factor analysis that justifies a unidimensional construct (Songer & Gotwals, 2012; Songer, Kelcey, & Gotwals, 2009; Gotwals & Songer, 2013; Gotwals & Songer, 2006b; Gotwals, Songer, & Bullard, 2012). The former strategy allows for separate appraisal of students' content knowledge and inquiry skills, while the latter presumes that content and inquiry cannot be separated in the context of their assessment items.

Examples of assessments that measure NGSS core ideas, practices, and crosscutting concepts.

Since the Next Generation Science Standards (NGSS) were released in 2013, a small number of projects have begun to develop assessments of science learning using the 3-dimensional structure prescribed therein. Two such efforts are described below. As both of the described projects are still in development, limited information is available about the methods that will be used to score and report student performance. However, the description of the constructs and items for assessment provide valuable information about the probable direction of future NGSS-aligned assessments.

Next Generation Science Assessment. Next Generation Science Assessment is a collaborative project among several assessment companies and universities with the goal of developing technology-enhanced assessment tasks aligned with the Next Generation Science Standards. The project specifically focuses on formative assessment for middle

school physical science and life science. They define the construct underlying the integration of core ideas, practices, and crosscutting concepts as "knowledge-in-use": the application of content knowledge in the context of disciplinary practice (DeBarger, et al., 2014). NGSS performance standards are broken down into smaller units called "learning performances," which form the basis for an assessment task. Learning performances are linked to all three assessment dimensions; however, a particular assessment task appears to be predominantly associated with only one NGSS dimension and implicitly linked to the remaining two dimensions (McElhaney, et al., 2015).

Multidimensional scaffolding. Many example items include multiple components – several constructed-response or multiple-choice items following the same stimulus material. These items appear to separate content and practice-specific skills into separate prompts (McElhaney, et al., 2015). For example, one middle-school science item presents the student with a table of data listing several unidentified substances and their properties. Students are asked to identify which two substances might be the same (content), and then provide an argument to support their choice (practice). As stated previously, however, each assessment task is predominantly linked with only one NGSS dimension. Thus, the purpose of using multiple components in an assessment task is unclear.

Response format. According to the stated goals of the program, some assessment tasks will involve interactive components (e.g., simulations, videos), however, an examination of currently available assessment items and curricular units reveals that they all use traditional paper-and-pencil item formats (constructed-response, multiple-choice), often in combination, i.e., a multiple-choice item with a constructed-response follow-up (see http://ngss-assessment.portal.concord.org/).

Since the project is still in development, the nature, motivation, and justification for some design decisions (e.g., the choice to use multiple-component items) is unclear. In particular, the alignment between the three NGSS dimensions and the item tasks remains ambiguous. It is unclear whether the items' underlying construct is an amalgamation of the three NGSS dimensions, one primary NGSS dimension with influence from the secondary dimensions, or if student responses will yield specific information about each dimension individually. Whatever the intended alignment, student responses need to be examined to ensure that the items reflect the construct.

Scoring and reporting. A description of the scoring process is not widely available because the tasks are still in an early stage of development. However, one scoring rubric for a task measuring the NGSS crosscutting concept "patterns" describes 6 possible levels of performance. The "patterns" task rubric explicitly emphasizes performance on the NGSS crosscutting concept, but it incorporates elements of the other two dimensions (core ideas and science practice) (McElhaney, et al., 2015). The task developers justify the use of a single rubric by emphasizing the importance of integrating, rather than separating, the three dimensions of science learning. They argue:

> The Framework and NGSS emphasize that CCCs, DCIs, and practices are necessarily and tightly intertwined. In authentic science, these dimensions do not occur in isolation of one another. As such, instruction and assessment need to integrate, rather than isolate, the three performance dimensions. A critically important quality of our rubric approach is that, despite foregrounding the CCCs,

> it acknowledges and preserves the rich connections that occur across the CCCs, DCIs, and practices. (p. 9)

It is unclear whether or not they intend for scores produced by their integrated rubrics to be aggregated as an indicator of overall "integrated" science learning, or if scores resulting from different rubrics will be grouped together according to the "foregrounded" dimension. The latter tactic would provide dimensional estimates that manage to simultaneously reflect the integration and distinction of the three dimensions, whereas the former reflects integration only and is more similar to the stance generally taken by large-scale science assessment systems.

NGSS Sample Classroom Assessment Tasks.

Multidimensional scaffolding. The NGSS released a set of sample classroom assessment tasks that draw on the NGSS and the Common Core State Standards for Mathematics (NGSS Lead States, 2014). These are lengthy, multi-component performance tasks designed to take up multiple class periods. Each task is linked to one or more performance expectations, with extensive descriptions justifying the use of individual task components to assess specific parts of the performance expectations, explaining which parts of the task comprise evidence of a particular standard, and describing how the three dimensions are integrated in the task. Within a task, an individual component requires the integration of multiple NGSS dimensions. This suggests that the components serve to break the task into manageable parts rather than to distinguish between NGSS dimensions within a task. The relationship between each task/task component and NGSS performance expectations/dimensions is clearly outlined.

However, these tasks are far too time-consuming for summative assessment – a single component often requires at least an entire class period for completion.

Response format. All classroom assessment tasks are open-ended, and each component of the items requires an extended response.

Scoring and reporting. Scoring rubrics for the NGSS sample assessment items are forthcoming, and will suggest scores for individual task components that describe levels of student performance as below basic, basic, proficient, and advanced. It is unclear how the suggested scoring rubrics account for the integration of three core ideas, practices, and crosscutting concepts in the assessment tasks.

Science Assessment Item Collaborative. The Science Assessment Item Collaborative (SAIC) is a joint effort between WestEd and the Council of Chief State School Officers, and includes an assessment framework, item specifications, and prototypes to guide development of summative assessment items for the NGSS. The item specifications and prototypes emphasize the *item cluster*, a group of items sharing a common context and performance expectation(s) as the main unit of assessment – their realization of the "multicomponent" item format suggested by Pellegrino, et al. (2014). Their effort is very closely aligned to the NGSS; each item cluster is mapped to a specific performance expectation(s), each item/item part is mapped to at least two of the NGSS dimensions, and each is linked to specific evidence statements from the NGSS (WestEd & CCSSO, 2015a; 2015b; 2015c).

Multidimensional scaffolding. In the SAIC item prototypes, scaffolding is used to link items together into clusters, and to splitting items into multiple parts (WestEd & CCSSO, 2015b). The SAIC Assessment Framework explicitly describes scaffolding as a

strategy to "guide students through a series of progressively more challenging interrelated questions to elicit evidence of what students know and are able to do." (WestEd & CCSSO, 2015a, p. 30) Within an item cluster, each item is aligned with at least two of the three NGSS dimensions. Scaffolding, therefore, is not a strategy to manage an item cluster's multidimensionality, but to enhance the items' accessibility to students.

Response format. The SAIC items utilize a combination of multiple-choice (and multiple-select), constructed-response, and various technology-enhanced item formats (e.g., drag-and-drop, drop-down menus, and graph modification items). Many different formats are used within an item cluster, and sometimes multiple formats are used within a single item. The item specifications guidelines include very detailed descriptions of the benefits and drawbacks of several different item formats, but do not include any discussion of multidimensionality (WestEd & CCSSO, 2015c).

Scoring and reporting. At this point in time, SAIC items are only prototypes, so there are no formal plans for scoring and reporting student responses. Some of the individual items contain notes about scoring information. According to these notes, each "item" within an item cluster receives a single score, despite the mapping of at least two NGSS dimensions to each "item" (WestEd & CCSSO, 2015b). This suggests that the item writers envision a unidimensional scoring method and measurement model.

Summary

Multidimensional scaffolding. Traditionally, multidimensional constructs have been measured by using separate, unconnected assessment tasks for each dimension (e.g., College Board, 2014c; ACT, 2016; Peoples, 2012; Gronmo, et al., 2013). In science, however, this approach seems unsuitable for the way that NGSS dimensionality is

defined as a set of integrated constructs. In response, many science assessment programs have opted to integrate content and practice within a single assessment task. The NAEP paper-and-pencil test, for example, classifies items according to one of three content domains and one of four science practices (National Assessment Governing Board, 2007). PISA items, similarly, assess one of three "competencies," one of three "knowledge types," and one of three "systems" simultaneously (OECD, 2012a; 2012b; 2015; 2017). AP science exam design emphasizes that every item incorporates content "Big Ideas" with practice in every item (The College Board, 2014a; 2014b; 2015). Although these strategies acknowledge the importance of both content and practice in the domain of science, specific information about student proficiency on the separate dimensions is unavailable.

One recommendation for providing unique information about student proficiency on multiple integrated dimensions of science learning is to use multi-component items (Pellegrino, et al., 2014). Some of the science assessments described here already use some form of a multi-component task design, usually by grouping together multiple selected-response or short-answer items (e.g., The College Board, 2015; Quellmalz, et al., 2012; Songer & Gotwals, 2012; McElhaney, et al., 2015). Such a design strategy seems like an intuitive way to integrate multiple dimensions into a single assessment task; however, in most cases the multi-component item structure did not directly map to the underlying dimensional structure as recommended in Pellegrino, et al (2014). Without the separation of dimensions among item prompts, it is unclear how multi-component items are able to enhance the multidimensional aspect of NGSS assessment.

Response format. It is evident that measuring multidimensional science proficiency elicits a range of strategies from assessment developers, especially in the way that the relationship between the constructs is defined and subsequently reported. However, there are many commonalities among these assessments; most notably, to a large extent they rely on traditional item types (multiple-choice/selected-response, constructed-response). Even the simulation-based assessments and computer-based tasks utilized multiple-choice or short constructed-response items to supplement informal data about student behavior within the simulation (Quellmalz, et al., 2012; OECD, 2015). These common item types, which have typically been used to assess a single competency, are now being extended to assessment of multiple competencies. In most cases it is unclear whether this design decision is justifiable, as assessment systems rarely make note of investigating the cognitive processes that are actually elicited by their items. On the whole, there is limited evidence to support the use of traditional item types to measure multidimensional constructs.

Scoring and reporting. The current status quo in science assessment is that student performance is scored and reported on one overall science scale, and content and practice are assumed to contribute equally to every student's performance (National Center for Education Statistics, 2009; Reshetar; 2012). This implies that the dimensions are indistinguishable from each other. While content and practice certainly influence each other throughout the learning process, conflating the two dimensions obfuscates the unique contribution of both content and practice to science learning. Information about

science content, practice, and crosscutting concepts, separately, may be valuable for science teachers, curriculum developers, researchers, and policymakers.

Some science assessments have assessed content and practice using integrated tasks, while still managing to retain the distinction between the 2 constructs. These tend to be smaller-scale assessments intended for classroom use. The BioKIDS project accomplished this by scoring student responses according to their content understanding and the quality of their scientific explanations. Even so, they report student data inconsistently, sometimes distinguishing between content and explanation (Gotwals, 2006; Gotwals & Songer, 2006a; Gotwals & Songer, 2010), and sometimes merging the two constructs together on a single metric of student performance (Songer & Gotwals, 2012; Songer, Kelcey, & Gotwals, 2009; Gotwals & Songer, 2013; Gotwals & Songer, 2006b; Gotwals, Songer, & Bullard, 2012). The SEPUP assessments (Roberts, Wilson, & Draney, 1997) also distinguish between content and practice by defining separate scoring categories for the content-related and practice-related elements of each assessment task. The SimScientist program (Quellmalz, et al., 2012) also administers integrated simulation-based tasks and reports information about students' content and practice separately. These three assessment projects share in common that they are focused on particular content domains and practices, rather than seeking to provide an overall assessment of student performance in science. The SEPUP assessment system and SimScientists also make an especially targeted effort towards making scores interpretable and useful for teachers so that they may target instruction in the dimensions where it is most needed. Based on these observations, it appears that distinguishable information

about content and practice is easier to obtain when the assessment domain is limited in scope.

Conclusion

This literature review has made it apparent that assessing multidimensional science constructs has long been a challenge for assessment developers. Many different strategies have been used to assess multidimensional science constructs, including different sources of information (e.g., simulation process data, assessment items), the size of the task (e.g., large multi-component task, single items, groups of related items), and response formats (selected-response, constructed-response). In addition, different scoring and reporting strategies have been used, with the result that some assessment results emphasize students' overall science proficiency while others reflect specific abilities related to smaller components of science proficiency.

Despite the apparent multitude of strategies for assessing science learning, there is little research supporting any particular item design decision (e.g., amount of scaffolding, response format) over another, especially when it comes to multidimensional constructs. Furthermore, investigation of the relationship between the assessment dimensions will provide information about the appropriateness of differentiating between student abilities on sub-dimensions of science performance. Such research is essential to ensure that interpretation of student assessment results is valid – a primary concern for assessment developers (American Educational Research Association, 1999).

Chapter 3: Methodology

The purpose of this dissertation is to explore the appropriateness of different item characteristics, scoring rubrics, and measurement models for a science assessment aligned with the three-dimensional Next Generation Science Standards. It will accomplish this by answering the specific research questions:

1. To what extent does multidimensional scaffolding affect the quality of information gained from students' responses to multidimensional assessment items?
    a. Does the impact of scaffolding vary for students of different abilities?
2. In the assessment of students' argumentation ability, does the use of a selected-response item format affect the extent to which the enacted construct reflects the intended construct?
3. To what extent do unidimensional and multidimensional scoring and modelling approaches affect the empirical relationships among the assessment items, and what does this imply about the relationships between the 3 dimensions of science learning?
4. How well does student performance reflect the hypothesized definitions of the underlying constructs and their relationships?

The rest of the chapter describes the path for exploring these research questions, including a) an overview of the data collection, b) the items design, c) background information about the qualitative and quantitative methods used for data collection and analysis, d) data collection procedures, e) the item revision process, f) scoring, and g) specific qualitative and quantitative procedures used to analyze the data.

Overview

To answer the research questions, information about the utility of multidimensional scaffolding (RQ 1) and selected-response formats (RQ 2) for multidimensional science assessment, and information about the empirical relationship between the three NGSS dimensions of science learning (RQ 3) was collected throughout the development process for an elementary science assessment. To investigate research questions 1 and 2, two rounds of cognitive interviews were conducted with 67 4th grade students from a public school district in the Boston metro area. The students responded to a small number of items and then reported about their experiences. Student responses and commentary about the items were qualitatively examined to see whether the amount of multidimensional scaffolding and/or response format had an impact on student response processes.

To investigate research questions 3 and 4, a pilot test was conducted with 369 4th,5th, and 6th grade students from school districts in the surrounding region. Student responses were scored according to two alternative rubrics – a holistic rubric and a multidimensional (analytic) rubric – and a psychometric analysis was conducted. Results provide information about the internal structure of the assessment in general, and more specifically about the impact of scoring decisions on the assessment's internal structure (RQ 3). Results also provide information about whether students' understanding of concepts related to matter, measurement, and argumentation align with the hypothesized construct definitions, and whether students' grade level and curriculum exposure are related to student understanding (RQ4). Data from the pilot test also allowed for empirical comparisons of different item variants (RQ 1 and RQ 2).

Throughout the research and development process, items were iteratively revised in accordance with research findings.

Items design

Research question 1 frames an investigation of item scaffolding in multidimensional assessment. In line with this research question, assessment items were manipulated so that three alternative versions of each item were created, all stemming from the same content and context but with the amount of *multidimensional scaffolding*[13] used in each assessment task varying slightly in each manipulation. The three variations are called *single-prompt explicit multidimensional*, *multiple-prompt explicit multidimensional*, and *single-prompt implicit multidimensional* items. A single-prompt explicit multidimensional version of an item contains the item stimulus (i.e., a text or visual description of the scenario under consideration) followed by a single response prompt. Although there is only one prompt, it explicitly asks students to attend to all three dimensions of science learning. The multiple-prompt explicit multidimensional version of the item has the same stimulus (i.e., identical content and setup of the assessment task), but the prompt is separated into multiple parts, with each sub-prompt intended to elicit a student response reflecting only one dimension at a time. Finally, the single-prompt implicit unidimensional version of an item is set within the same context as the single-prompt explicit multidimensional and multiple-prompt explicit multidimensional versions of the item, but with a single task that draws upon all three dimensions of science learning without explicitly asking students to attend to all three dimensions. The creation of explicit and implicit versions of an item allows for

[13] A definition of multidimensional scaffolding, and justification for use of the term *scaffolding* in assessment can be found in Chapter 2.

comparison of performance when students are faced with the same task embedded among different amounts of task scaffolding (single-prompt explicit multidimensional vs. multiple-prompt explicit multidimensional items vs. single-prompt implicit multidimensional items). Maintaining identical content and task context across the different versions of an item ensures that differences in item content and task are not confounded with differences in the amount of multidimensional scaffolding, so that variation in student performance is attributable solely to variation in the amount of scaffolding. An example item with three alternative versions, each with a different amount of multidimensional scaffolding, can be found in Figure 3.1.

Research question 2 poses an examination of selected-response and open-ended item response formats and their effect on student response processes. To address this question, two of the three scaffolding variations (multiple-prompt explicit multidimensional and single-prompt implicit multidimensional) were modified to include selected response components. These selected-response variations underwent a second round of cognitive interviews.

Data Collection and Analysis Methods

The purpose of research questions 1 and 2 is to determine how different item characteristics (i.e., multidimensional scaffolding and response format) affect the alignment between the intended and actual construct of multidimensional science assessment items. This is an essential part of an instrument's validity (Pellegrino, Chudowsky, & Glaser, 2001), and therefore is an important topic of research.

Researchers have successfully used a wide variety of methods to uncover the construct accessed by respondents to an assessment item, usually by examining responses

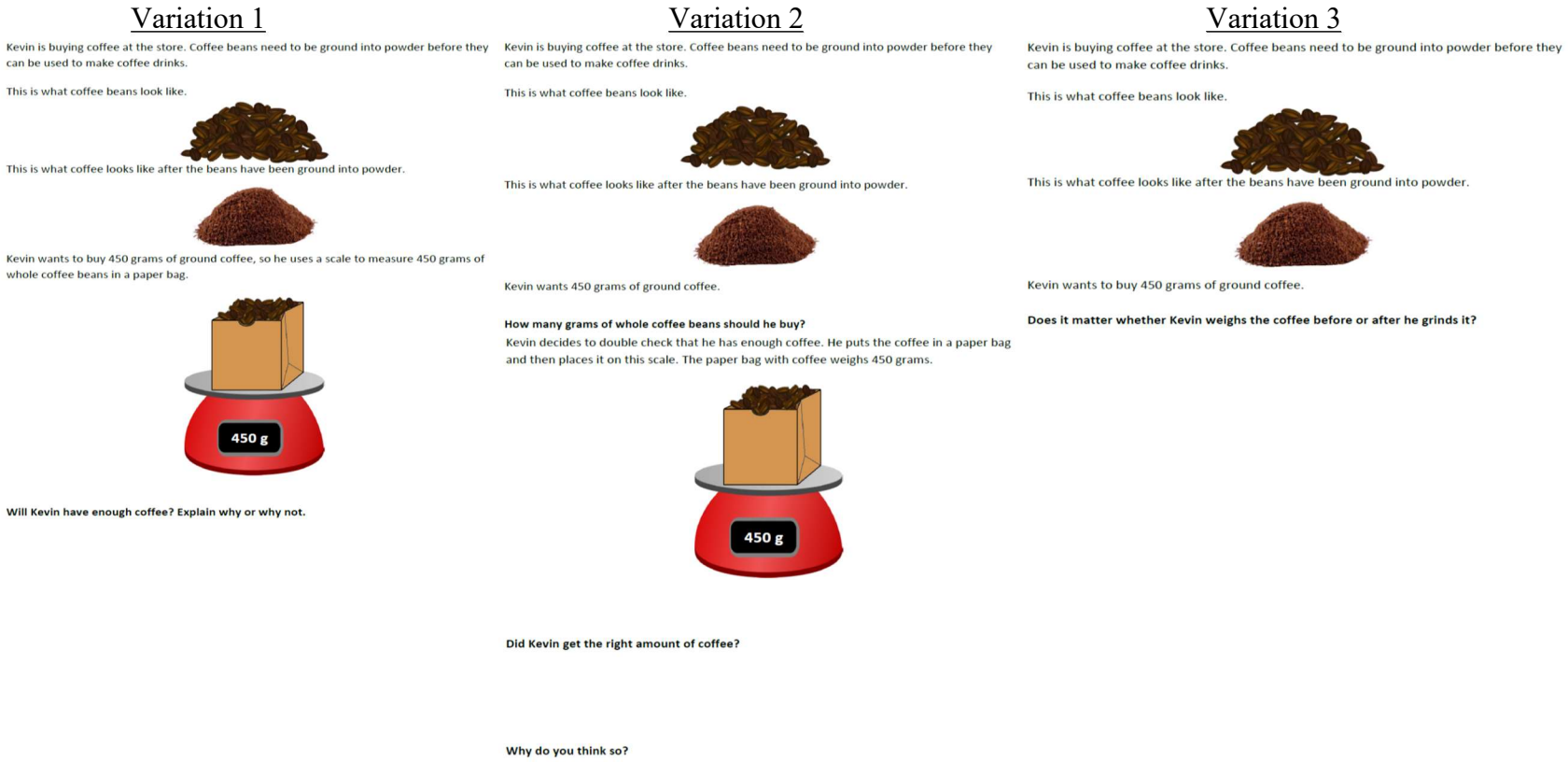


Figure 3.1. Three variations of an item with different amounts of multidimensional scaffolding. Variation 1 is a single-prompt explicit multidimensional item. In this item, a single prompt captures information about all three dimensions: Structure of Matter, Engaging in Argument from Evidence, and Scale, Proportion, and Quantity. Variation 2 is a multiple-prompt explicit multidimensional item. In this item, the first prompt captures information about the Structure of Matter dimension, the second prompt captures information about the Scale, Proportion, and Quantity dimension, and the final prompt captures information about Engaging in Argument from Evidence. Variation 3 is a single-prompt implicit multidimensional item. In this item, the prompt captures information about only one dimension: Structure of Matter.

and response processes. The most commonly used methods include cognitive interviews, observation of response time, and eye-tracking (American Educational Research Association, 1999; Wilson, 2005; Gorin, 2006). Statistical methods may also be used to link particular item features (e.g., response format, reading demand, etc.) to indicators of the enacted construct (e.g., item difficulty estimates) using an approach called item difficulty modeling (Gorin, 2006). Experimental designs can be used to examine the causal impact of item features on the enacted construct by creating multiple versions of an item, manipulating one or more of the item's features (format, language, context, etc.), and examining changes in observable characteristics of the responses (usually parameter estimates, but also response time) (Enright, Morley, & Sheehan, 2002; Gorin, 2005; Katz & Lautenschlager, 2001).

This study compared several variations of elementary science assessment items by using cognitive interviews to investigate differences between the intended and the enacted construct via direct comparisons of the rich data gleaned from student reports. Cognitive interview data was collected about three variations of multidimensional scaffolding (single-prompt explicit multidimensional, multiple-prompt explicit multidimensional, and single-prompt implicit multidimensional) and two response format variations (selected-response and constructed-response). Results provide information about the item types' relative appropriateness for use in measuring three-dimensional science learning.

The purpose of research question 3 is to investigate the assessment's internal structure. Internal structure is another critical source of validity evidence, as it confirms whether the items conform to the assumed structure of the construct (American

Educational Research Association, 1999). The structure of the construct has implications for the interpretation and statistical summarization of student responses, and is a particularly important consideration for NGSS assessment because the proposed assessment framework explicitly outlines three sub-domains of science learning (NGSS Lead States, 2013; Committee on a Conceptual Framework for the New K-12 Science Education Standards, 2011). The NGSS definition of the construct thus implies that student performance should be reported and interpreted with respect to all three sub-domains. According to the AERA Standards for Educational and Psychological Testing (1999), "when interpretation of subscores…is suggested, the rationale and relevant evidence in support of such interpretation should be provided." (p. 20). This study produces evidence for the interpretability of core idea, practice, and crosscutting concept subscores by establishing the dimensional structure and subsequent technical quality of the resulting assessment dimensions.

There are several common ways to establish or confirm the internal structure of assessment data. Under the principles of classical test theory, factor analysis can be used to reveal the number of factors (or underlying latent groupings of items) by examining the patterns of covariance among a group of items (e.g., Stone, Ye, Zhu, & Lane, 2010). Alternatively, software packages like DIMTEST (Stout, 1987), DETECT (Zhang & Stout, 1999), and NOHARM (Fraser & McDonald, 2003) each utilize a specific method and statistic to test competing dimensionality structures. In this study, the Rasch family of item response theory (IRT) models will be used to explore the internal structure of an NGSS-aligned science assessment by comparing the fit of unidimensional and multidimensional models (Adams, Wilson, & Wang, 1997).

Cognitive interviewing. Cognitive interviewing has its roots in cognitive psychology, where it originated as a way to investigate the cognitive processes underlying behavior (Ericsson & Simon, 1993). The field of psychology is fraught with debate about the believability of a person's subjective report of their own cognitive state (Nisbett & Wilson, 1977; Verplanck, 1962), but if we operate under the assumption that subject's verbal reports can be trusted– a viable assumption (Ericsson & Simon, 1993) – then the cognitive interview can be used as a tool to investigate a great variety of research problems. One application of cognitive interviewing techniques is in the field of measurement, where it can be used to improve the quality of survey instruments (Willis, 1999; Beatty & Willis, 2007; Drennan, 2003; Knafl, et al., 2007; Ryan, Slater, & Culbertson, 2012). In educational measurement, in particular, the cognitive interview has been used with great success to support the development and validation of knowledge and attitudinal assessments (Almond et al., 2009; Demisone & Le Floch, 2004; Howell, Phelps, Croft, Kirui, & Gitomer, 2013).

Originally, cognitive interviewing was dominated by a single technique: the think-aloud. In a think aloud interview, participants are prompted to recount their complete thought process as they encounter and respond to an item (Ericsson & Simon, 1993). The role of the interviewer is to facilitate the examinee's verbal report without directing it – after instructing the interview participant about how to think aloud, the interviewer remains unobtrusive, at most offering a reminder for the participant to continue verbalizing their thoughts (Beatty & Willis, 2007). The advantages of this method are several: the interviewer's minimal role provides little opportunity to introduce bias, the method is simple enough that interviewers do not need to be extensively trained,

the open-ended nature of the method leads to a possibility of unanticipated information from the interviewee, and the interviewee's verbalization occurs concurrently as they experience the item which makes their stream of consciousness a more "pure" data source than a retroactive report (Willis, 1999; Beatty & Willis, 2007). However, there are some drawbacks to the think-aloud method, most notably that participants often find it difficult to provide the level of detail that is most useful for researchers, the interview subject is in control and thus might divert their attention to processes that are irrelevant or not useful, and that the extra effort of verbalizing might constrain or enhance the amount of cognitive processing experienced by the interviewee (Willis, 1999; Beatty & Willis, 2007).

Other cognitive interviewing methods have emerged and become widespread, including various types of probing and even unique tactics like role-play (Willis, 1999; Van Someran, Barnard, & Sandberg, 1994). Probing differs from the think-aloud in that the interviewer has greater control of the structure and content of the interview, and can interject questions about the interviewee's comprehension, confidence, and response strategies, respond to observed student behaviors, clarify the participants' meaning, or other specific issues of interest (Willis, 1999). Although this provides information that is much more relevant to the researcher's specific concerns, it potentially introduces additional elements of bias and artificiality into the interview by interrupting the flow or redirecting the content of the interviewee's thoughts. Thus, an important consideration when developing interview probes is specificity: probes should be broad enough that they do not direct the interviewee towards a particular answer, but specific enough that they produce information that is of interest to the interviewer (Willis, 1999; Beatty & Willis,

2007). Probes may be scripted, or developed prior to the interview and standardized, or spontaneous. Scripted probes allow for better comparability between respondents, but spontaneous probes may potentially provide richer information or unearth unanticipated issues (Willis, 1999; Beatty & Willis, 2007). Beatty and Willis (2007) suggest that both types of probes may sometimes be appropriate during a single interview. Additionally, probes may occur concurrently as the student encounters the item or retrospectively, after all items have been encountered. Concurrent probes may be more desirable because they request information that may still be easily accessible in short term memory, whereas retrospective probes may result in generalization or inference about a series of experiences (Ericsson & Simon, 1993). However, some research suggests that retrospective probing may elicit more statements about the chosen response, which may be of interest depending on the specific research question (Kuusela & Paul, 2000; Howell, et al., 2013).

Cognitive interviews result in a record of examinee statements about their experiences and opinions related to an assessment item, rather than numerical data. Therefore, cognitive interview data is commonly analyzed by qualitative coding of examinee statements (Almond et al., 2009; Howell et al., 2013). Howell, et al. (2013) developed a coding scheme centered around specific research questions related to the respondent's reasoning and opinion on the realism of items. The examinees' responses to each probe were coded according to a number of different criteria. The coding categories corresponded to the potential anticipated content of the response. The resulting codes were then summarized using descriptive statistics. Development of a coding scheme with different criteria and categories of anticipated responses is recommended before the

interviews are conducted, and is especially useful when a standardized interview protocol is used (Almond et al., 2009). Inductive categorization, the use of categories that emerge from the data rather than an a priori categorization scheme, may also be useful to account for unanticipated issues and/or responses to spontaneous probes (Rossman & Rallis, 2012).

In this study, the cognitive interviews utilized probing rather than think-aloud techniques (Beatty & Willis, 2007; Ericsson & Simon, 1993) because probing provides more structure and direction for verbal feedback, which is recommended when the interview sample has poor metacognitive or verbal skills (Almond et al., 2009) – a plausible concern for the sample of 4th grade students. The interview protocol contained a series of standardized questions related to the research questions and other anticipated issues of concern. Students were asked how they understood the task required by each response prompt, to explain the rationale for their given response, and potentially about the meaning of specific terms included in their response. Spontaneous probes were also used when warranted to clarify a participant's response or reaction to an item. A set of standardized prompts ensured that each child was given the opportunity to provide information about all of the criteria that were used to evaluate the different item characteristics, and spontaneous probing brought attention to unanticipated issues.

The Rasch family of item response models. Item response theory (IRT) models are probabilistic models that represent the probability of an item response as a function of the respondents' underlying ability and the items' characteristics (e.g., difficulty, discrimination, and guessing) (de Ayala, 2009), and are seen as useful tools for educational measurement. The Rasch family of models is a subset of IRT models that

mainly focus on only one item characteristic: difficulty. Rasch estimates of item difficulty and person ability are on an equal interval scale, meaning that an equal difference between any pairs of estimates has the same meaning with regard to the underlying construct, no matter where the estimates fall on the overall scale (Boone & Scantlebury, 2006). Rasch estimates are also considered to be sample and test independent: item and person ability estimates are stable regardless of the particular items on the test or characteristics of the respondents, within limits (Boone & Scantlebury, 2006). This is vastly different from classical test theory, where "harder" and "easier" tests do not provide comparable estimates for students, and estimates of an item or test's difficulty depend on the relative ability of the sample (Lord, 1980). The Rasch family of models are particularly useful because of their strong inherent link to a carefully defined construct (Wilson, 2005). Values of person estimates reflect the amount of examinees' ability related to the underlying construct. Similarly, item estimates indicate which tasks require examinees to demonstrate more or less ability on the underlying construct. Item and person estimates can be placed on a scale, with their relative position providing empirical information about the progression of ability and behavior related to the underlying latent construct. The Wright map is a visualization of this information, with person and item estimates located on opposite sides of a vertical continuum, which represents the underlying construct (Wilson, 2005). Item and person estimates are arranged vertically from least to greatest, so that estimates at the bottom of the figure represent students with less sophisticated understandings and easier items. The estimates at the top of the figure represent more sophisticated understandings and difficult items. In

addition, Rasch models can accommodate complex constructs that include multiple dimensions of a latent ability (Adams, Wilson, & Wang, 1997).

The basic Rasch model gives the probability of a dichotomous score (e.g., correct or incorrect) as a function of the difference between the respondent's *ability* and the item's *difficulty* (Rasch, 1960; 1980). The model takes the statistical form:

$$P(X_{ni} = 1|\theta_n, \delta_i) = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)}$$

where X_{ni} is the score of a response given by person *n* with ability θ_n on item *i* with difficulty parameter δ_i. According to the model, large positive differences between person ability θ_n and item difficulty δ_i (i.e., relatively easy items and/or relatively able examinees) correspond to a high probability of a correct response. The converse is true when the difference between θ_n and δ_i is negative (i.e., relatively low ability examinees and/or relatively difficult items). When an item's difficulty matches a person's ability (i.e., $\theta_n = \delta_i$), the probability of a correct response is 50% $\left(\frac{\exp(0)}{1+\exp(0)} = \frac{1}{2}\right)$.

The model can be extended to include scores that reflect intermediate levels of correctness. This is called the Partial Credit Model (Masters, 1982), and takes the following form.

$$P(X_{ni} = x_i|\theta_n, \delta_{ik}) = \frac{\exp \sum_{k=0}^{x_i}(\theta_n - \delta_{ik})}{\sum_{j=0}^{m_i} \exp[\sum_{k=0}^{j}(\theta_n - \delta_{ik})]}$$

Here δ_{ik} is the difficulty parameter associated with step *k* of *m* possible steps for item *i*. A "step" refers to the increase in item difficulty associated with increasing partial credit scores. (For model identification, $\sum_{k=0}^{0}(\theta_n - \delta_{ik}) \equiv 0$.) As in the dichotomous model, greater "ability" corresponds to a higher probability of achieving a larger score on an item (i.e., surpassing more steps).

Multidimensional constructs. Dichotomous and polytomous Rasch models have been used extensively in the assessment of educational achievement (Boone & Scantlebury, 2005; Wilson & Sloane, 2000), and in surveys of student attitude/affect (Boone, Townsend, & Staver, 2010; Martin & Mullis, 2012). These forms of the model assume that only one latent construct (e.g., "mathematics ability", or "self-efficacy") influences responses, an assumption commonly referred to as unidimensionality (Rasch, 1960; 1980; Ludlow, et al., 2014). All item responses are considered probabilistic manifestations of this latent ability. Consequently, items are placed on a single scale. However, in certain situations multiple related abilities might contribute to a response, either directly or indirectly. For example, using the framework for science learning employed by the NGSS, a student's conceptual understanding of a topic (*core idea*) might reinforce their interpretation of data in an assessment activity measuring a science *practice*. Or, an assessment task might require the student to construct a scientific argument that requires both conceptual knowledge (*core idea*) and the interpretation of a mathematical model (*practice*). In these cases, the use of a unidimensional model is a "theoretical impurity", presenting an overly simplified view of student performance and conflating the relationships among multiple latent abilities (Adams, Wilson, & Wang, 1997, p. 11). Furthermore, the influence of additional unaccounted-for constructs can negatively impact the technical quality of scale estimates. If multiple dimensions underlie a set of assessment items, and the correlation between dimensions is not high, then the use of a unidimensional model can bias the item and person estimates (Folk & Green, 1989; Ackerman, 1992).

The MRCMLM. Another extension of the Rasch model, the Multidimensional Random Coefficients Multinomial Logit Model (Adams, Wilson, & Wang, 1997) allows for violations of unidimensionality by including multiple latent variables to explain student responses. In this case, the student ability parameter (θ_n) and item difficulty parameter (δ_i) from the unidimensional Rasch model are expanded to vectors to include additional latent abilities that may also influence item responses.

$$P(X_{ik} = 1, \boldsymbol{A}, \mathbf{B}, \xi | \boldsymbol{\theta}) = \frac{e^{\mathbf{b}_{ik}\boldsymbol{\theta} + \mathbf{a}_{ik}'\xi}}{\sum_{k=1}^{K_i} e^{\mathbf{b}_{ik}\boldsymbol{\theta} + \mathbf{a}_{ik}'\xi}}$$

In this formulation, the student ability parameter θ_n becomes a column vector with *D* rows corresponding to the number of dimensions represented in the model. The item and step parameters are in vector ξ. The scoring matrix **B** assigns values (scores) and dimensions to particular item responses, and the design matrix **A** assigns parameters to items and response categories. (The scoring and design matrices are not model parameters, but are critical for model specification because they map items and scores to dimensions.) Thus, student responses to each item are directly or indirectly explained by multiple latent variables. Determining the probability of a correct response (or a specific category response in a partial credit item) is no longer as straightforward as the difference between ability and difficulty, but now is dependent on multiple ability parameters and the relationships between them, in addition to the differences between the ability parameters and item difficulty parameters. Thus, two examinees with the same ability θ_{n1} on dimension 1 may have different response probabilities on an item measuring dimension 1 if they have different abilities on the other, related dimensions.

Although multidimensional Rasch models are not as commonly used in educational research as unidimensional variations of the Rasch model, the MRCMLM

has been successfully used on assessments covering multiple related topics or skills (Briggs & Wilson, 2003; Liu, Wilson, & Paek, 2008; Paek, Peres, & Wilson, 2009). In addition to mitigating the theoretical and technical drawbacks associated with using a unidimensional analysis when assessment data has a more complex underlying structure, the MRCMLM also provides valuable refined information about the examinees' specific patterns of strength and weakness (Briggs & Wilson, 2003; Liu, Wilson, & Paek, 2008), provides unattenuated estimates of the correlation between dimensions (Briggs & Wilson, 2003), and produces ability estimates that have high utility as subscores (Dwyer, Broughton, Yao, Steffen, & Lewis, 2006).

ConQuest software (Adams, Wu, & Wilson, 2015), version 4, was used to estimate the MRCMLM. A marginal maximum likelihood (MML) estimation method was used. In contrast to a joint maximum likelihood (JML) estimation method, in which the person and item parameters are estimated concurrently, MML estimation means that the item parameters are estimated first under an assumed normal population distribution of ability (de Ayala, 2009). If the examinee population is non-normal, then estimates may not be representative of their true values; however, in general MML estimation produces parameter estimates with less bias than JML estimation, especially for tests with a small number of items.

NGSS assessment dimensionality. To determine the dimensionality of an NGSS-aligned science assessment, student response data from two alternative scoring rubrics was analyzed using unidimensional and multidimensional Rasch models. Model-data fit was compared across the proposed scoring rubrics and dimensional structures. Presumably, the best-fitting model is the one most consistent with the internal structure of

the assessment; therefore, results of this analysis tell us about the relationships among the three specific aspects of science learning assessed (Structure of Matter; Engaging in Argument from Evidence; and Scale, Proportion, and Quantity), and the effect that scoring decisions have on this relationship. The accuracy of the comparison of different dimensional structures based on model fit is high (> 95%) for sample sizes larger than 100, even when the correlations between dimensions are large (> 0.75) (Harrell & Wolfe, 2009). In the present study, the overall sample size is N = 369, and the sample sizes for individual items range from N = 92 to N = 310. Therefore, it is likely that model comparisons will provide an accurate identification of the most appropriate dimensional structure.

The unidimensional and multidimensional frameworks imply different theoretical structures for the underlying construct. Under a unidimensional interpretation, student responses on assessment tasks related to a core idea, practice, and crosscutting concept may be explained by the same underlying latent trait (e.g., overall science ability). This means that items measuring all three aspects of science learning can be analyzed together on a single scale without misrepresentation or loss of information related to additional explanatory latent variables. A unidimensional analysis is the most parsimonious, and thus is the most desirable when supported by the data. If the data do not support the unidimensionality assumption, then item and person parameter estimates will be biased (Folk & Green, 1989).

A multidimensional approach, on the other hand, implies that the measured core idea, practice, and crosscutting concept are not explained by the same underlying latent trait, but are actually separate, related constructs, each of which directly or indirectly

contributes to student performance on science assessment tasks. A multidimensional Rasch scale can be estimated consecutively or simultaneously. Using a consecutive approach (Davey & Hirsch, 1991), multiple subscales are acknowledged, but the subscales are assumed to be uncorrelated and parameters for each dimension are estimated separately. If simultaneous estimation is used, the multidimensional Rasch model both accounts for and utilizes covariation between multiple dimensions when producing person and item estimates (Briggs & Wilson, 2003). In this study, simultaneous estimation will be used, as it reduces the error in person estimates when dimensions are correlated, compared to a consecutive approach (Adams, Wilson, & Wang, 1997).

The NRC framework and the NGSS specify three dimensions which correspond to distinct entities, with clear and distinguishable definitions and separate sets of learning progressions, but they also describe a vision for science education where "students…actively engage in scientific and engineering practices and apply crosscutting concepts to deepen their understanding of the core ideas in these fields." (Committee on a Conceptual Framework for New K-12 Science Standards, 2011, p. 9). Thus, although they are defined separately, the three dimensions of science learning "must be woven together." (p. 29). This description implies that the writers of the NRC framework and the NGSS conceived of a multidimensional relationship among core ideas, science and engineering practices, and crosscutting concepts, such that the three dimensions are separate but related entities. This implication is further supported by language employed in the standards themselves, which explicitly refer to *three dimensional learning* (NGSS Lead States, 2013), and this language is repeated in many subsequent publications (e.g.,

Pellegrino, et al., 2014). Therefore, the use of a multidimensional measurement model is supported by the description of the structure of the NRC framework and NGSS. This study investigates whether this three-dimensional structure is also empirically supported by the results of an elementary science assessment about matter.

Data collection procedures

Cognitive interviews – 1st round.

Sample. To investigate the effect of multidimensional scaffolding on student response processes, 30 items underwent a cognitive interview process with 26 fourth-grade students from a suburb of Boston. Students were volunteers from 4 elementary classrooms in 2 elementary schools, and were recruited by their classroom teacher. All cognitive interview participants were current or previous participants in the fourth-grade Inquiry Project curriculum.

The school district was chosen for participation in this study because of their affiliation with a larger NSF-funded project[14] at TERC, which this research supported. During the 2015-2016 academic year, 8.2% of students were classified as economically disadvantaged, compared to 27.4% of students statewide. The district outperformed the overall Massachusetts proficiency rate for all subjects and grade levels on the 2015 state assessments (Massachusetts Department of Elementary and Secondary Education, 2016). Therefore, it is unlikely that the students who participated in this study are representative of the broader population of elementary students in the United States.

[14] That project, VideoReView, is the collaborative work of TERC, intuVision, Boston College, and Boston area teachers and supported by funding from the National Science Foundation, grant #1415898. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Items. In order to minimize the amount of time required from each student, the students encountered only 5 items throughout the course of an interview. Items were distributed across students such that each of the 30 items was used in at least 4 student interviews. Each student encountered at least one item from all of the possible multidimensional scaffolding variations (i.e., at least one single-prompt explicit multidimensional item, one multiple-prompt explicit multidimensional item, and one single-prompt implicit multidimensional item). Furthermore, students encountered a maximum of one version of an item from a particular task context to maintain independence of observations and eliminate the possibility of a learning effect. Item presentation was counterbalanced across students so that the same item content and/or multidimensional scaffolding variation did not consistently lead or conclude an interview session.

Procedure. Cognitive interviews took place during the students' regular class time. Interviews were conducted individually in a quiet location proximal to the students' classroom. Each student encountered one item at a time. First, the student was given the opportunity to silently read an item and provide a written response. Students were asked to notify the researcher when they finished responding. After they completed their response, the verbal interview began. The interview protocol can be found in Appendix A. The researcher asked the student a series of standardized questions about their understanding of the assessment task, the language of the item, and the reasoning underlying their response. If a student's verbal or written response was at any point unclear or surprising, the researcher followed up with additional, unscripted questions asking the student to elaborate. Once the student finished the verbal interview for an item,

they were given the next item and asked to silently read and provide a written response. The written and verbal portions of the interview proceeded iteratively until all 5 items had been administered. This took about 30-40 minutes per student.

The researcher recorded the amount of time required by the student to complete each item, from when the student first began reading to when they signaled that they had completed their written response. Student verbal responses were recorded with an audio recorder on the researcher's mobile phone. Audio files were initially saved to the phone's internal storage, then moved to a secure server and deleted from the phone. Student written responses to the assessment items were also collected for analysis and destroyed after being scanned to a secure server.

Cognitive interviews – round 2.

Sample. To investigate the effect of response format on student response processes, 28 additional items underwent a cognitive interview process with 37 additional fourth-grade students. Students were volunteers from 3 elementary classrooms in 3 elementary schools in the same school district, and were recruited by their classroom teacher. One of the schools that participated in Round 1 of the cognitive interviews also participated in Round 2, but the sample of students participating in the first and second rounds of cognitive interviews did not overlap. Again, all participants were current or previous participants in the fourth-grade Inquiry Project curriculum.

Items. Twenty-eight items were tested. The items were distributed across students as in the first round, such that students encountered a variety of levels of multidimensional scaffolding and response format in different item contexts, and the order of presentation was counterbalanced across students. The items used in this round

included a subset of the 30 items used in the previous round of cognitive interviews (revised based on student feedback), and corresponding selected-response versions of these items. During the cognitive interviews, students who saw the selected-response variation of a multiple-prompt multidimensional item were asked to complete an open-ended argument prompt followed by a selected-response version of the same prompt. Students who saw the selected-response variation of a single-prompt implicit multidimensional item were asked to complete an open-ended version of the prompt followed by a selected-response version of the same prompt, and then another prompt which explicitly cued the student to provide a written argument. The final prompt was only administered if the student did not provide a written argument as part of the first open-ended prompt. For both selected-response item variations, students were not allowed to go back and change an answer once they saw the response options, and any attempt or desire to change a previous answer was recorded. By asking students to complete both open-ended and selected-response versions of the same prompts, a direct comparison between the response formats was enabled.

Procedure. The cognitive interview procedure was largely identical to the procedure used in the first round of cognitive interviews. The interview questions focused on additional issues of concern, including some related to the response formats (e.g. why they picked one answer over the others, whether they understand any new words/phrases, or the meaning of distractors).

Item revision and selection for pilot test. Items underwent revision at two stages in the research and development process. After the first round of cognitive interviews, interviews were transcribed and coded to reflect areas of misunderstanding or

confusion[15]. Subsequently, the items were revised to address any issues that were uncovered. For example, where student responses indicated that vocabulary was difficult to understand, item language was changed to clarify confusing terms or phrases. In extreme cases, tasks that did not provoke productive student responses were replaced or heavily revised. At this point, an additional selected-response version of each item was created for examination in the second round of cognitive interviews.

Items underwent another round of revisions based on the results of the second round of cognitive interviews. This resulted in further changes to item language, revision of confusing or misinterpreted tasks, and replacement of item tasks that did not elicit informative responses. These revisions are important to the test development process: an iterative process of item development and revision improves the overall quality of all items, and ultimately leads to better test validity.

Finally, student verbal and written responses from the first and second round of cognitive interviews informed the number and type of item variants included on the pilot test. Based on student responses to the cognitive labs, it was clear that items with explicit scaffolding resulted in more complete, evaluable student responses[16]. Therefore, the single-prompt implicit multidimensional item variant was excluded from the pilot test. A subset of the remaining item scenario and variant combinations were chosen for the pilot test, including some multiple-prompt explicit multidimensional items and some single-prompt explicit multidimensional items, as well as some open response and some selected-response variants.

[15] For more details on the coding and transcription process, see the Analysis Methods for research question 1.

[16] For more details, see results of the first round of cognitive interviews in Chapter 3.

After item revisions were complete, one additional item scenario was created to fill in a perceived gap in the remaining items' content coverage. Two versions of this scenario (items 9A and 9B) were included in the pilot test, even though these items did not go through the cognitive interview process described above.

Pilot test.

Instrument. After the items were revised based on student feedback from the cognitive interviews, 11 item scenarios remained. Each scenario was used as the basis for 1-2 item variants, resulting in 20 items. There were three types of item variants: open-ended multiple-prompt explicit multidimensional items, selected-response multiple-prompt explicit multidimensional items, and open-ended single-prompt explicit multidimensional items. Four scenarios were used as a common basis to compare response format, by holding the scaffolding level constant (multiple-prompt explicit multidimensional) and varying the response format (constructed-response or selected-response). Five scenarios were used as a common basis to compare scaffolding levels, by holding the response format constant (constructed-response) and varying the amount of multidimensional scaffolding (multiple-prompt explicit multidimensional or single-prompt explicit multidimensional). See Appendix F for all 20 item variations used in the pilot test. The 20 items were distributed across 5 test forms, with each test form containing 9-10 multidimensional items. It was estimated that most 4th grade students should be able to complete 9-10 items within a 45-minute class period, and teacher reports confirmed this, although in most classrooms a few students were given extra time to complete the test. Each test form included items with different amounts of scaffolding and different response formats, to facilitate the direct comparison of item function across

item variants. Two items (items 10 and 11) were held constant and included on all 5 test forms as linking items. Linking the test forms enabled the direct comparison of item variations from a common scenario. Separate variations of the remaining items were administered on different test forms, so that psychometric indicators of item function could be directly compared between pairs of items that share identical content/task context but with different response formats or amounts of multidimensional scaffolding.

A matrix sampling approach was used to construct the test forms (Mislevy, Beaton, Kaplan, & Sheehan, 1992). Matrix sampling enhances the analysis of item and student performance by ensuring that all items can be scaled concurrently, regardless of whether they were taken by the same students. Each test form shared some items with each of the other test forms, including two items which were included on all forms. By using the common items as anchors, item and person estimates from multiple forms were concurrently calibrated on the same scale so that they are directly comparable. Although this method provides less precise estimates of student ability than an assessment where all students take the same test form, it allowed for the inclusion of a greater number of items on the pilot test. The distribution of scenarios and item variations across test forms can be found in Figure 3.2.

Sample. The pilot test was administered to a sample of 369 fourth, fifth, and sixth grade students from public schools in Massachusetts and Vermont. According to Linacre's recommendations for stable item calibration (1994), to obtain an item estimate with a 95% confidence interval within +/- ½ logit, sample size for each item should fall somewhere between 64 and 144 students. Because there were 5 evenly distributed test forms, each item was seen by ~148 students, which exceeds Linacre's recommendation.

Information about sample composition with regard to grade level, school district, and matter-related educational experiences can be found in Table 3.1. Recruiting students with a variety of ages and previous matter-related educational experiences was a priority, as it provides information about the range of student performance at varying levels of knowledge/ability.

Teachers were recruited for the sample by a state or district science administrator. As a small incentive for participation in the pilot test, teachers received feedback on their students' performance and the research findings. Individual student responses were not identified to teachers, but instead the feedback focused on student performance in the aggregate.

Procedure. The final test forms were administered to students during the time normally reserved for their science class. Teachers were asked to rotate all 5 test forms in their classrooms. Classroom teachers were responsible for test administration, and students were given at least one full class period (approximately 45 minutes) to answer all questions. Tests were administered in a paper-and-pencil format, and students were asked to respond directly on the test form. For any students who failed to complete all items within a single class period, the provision of extra time was at the discretion of the teacher. Teachers were asked to make a note of any issues encountered during test administration, including unusual disruptions, student questions about test items, extra time, and any other relevant concerns.

Background characteristics. At the classroom level, information was collected about grade level, and whether the class has participated in the Inquiry Project

Scenario	Item	Item Variation			Test Form				
		1	2	3	1	2	3	4	5
1	1A	X			X	X			
	1B			X			X		X
2	2A	X			X	X			
	2B		X				X		X
3	3A	X				X			X
	3B			X	X			X	
4	4A	X				X			X
	4B		X		X			X	
5	5A	X			X		X		
	5B		X			X		X	
6	6A	X			X		X		
	6B			X		X		X	
7	7A	X					X	X	
	7B			X	X				X
8	8A	X					X	X	
	8B			X	X				X
9	9A	X						X	X
	9B		X			X	X		
10	10	X			X	X	X	X	X
11	11	X			X	X	X	X	X
Total Number of Items	20	11	4	5	10	9	9	9	9

Figure 3.2. Distribution of item scenarios and variations across test forms. Item variations 1, 2, and 3 refer to open-ended multiple-prompt explicit multidimensional, selected-response multiple prompt explicit multidimensional, and open-ended single-prompt explicit multidimensional, respectively.

Table 3.1.

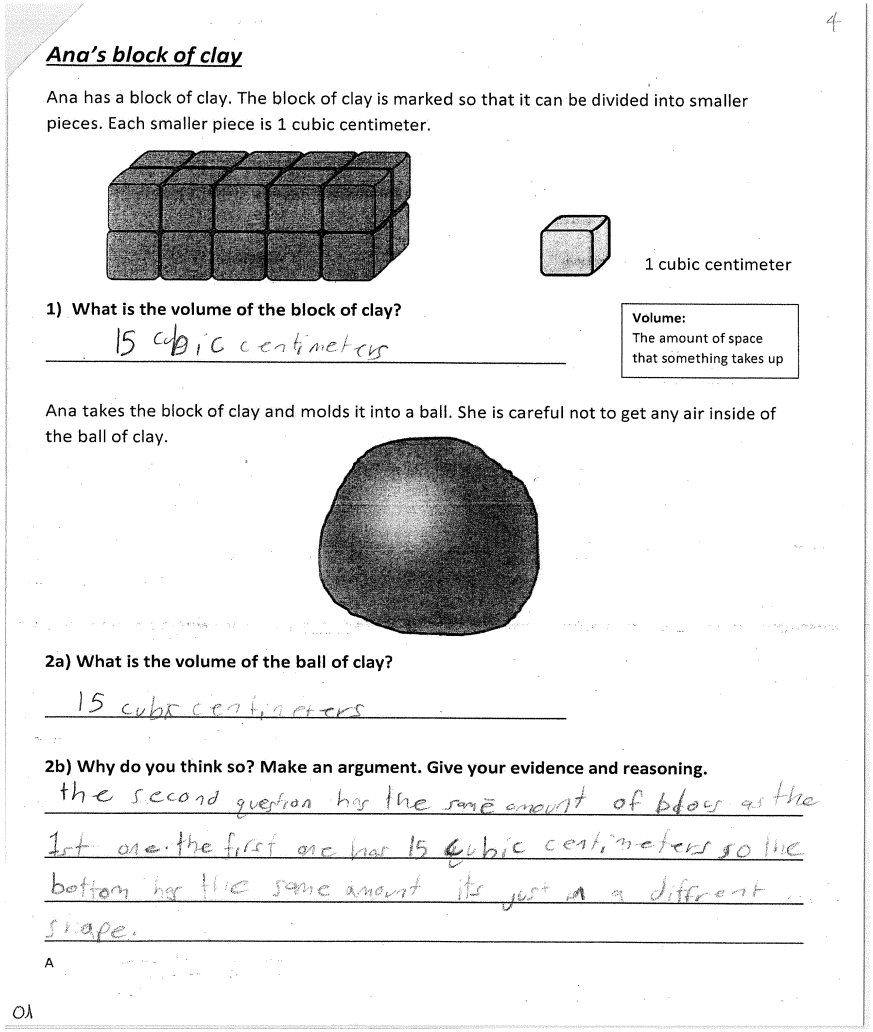
Sample Composition

Grade Level	Massachusetts	Vermont
Grade 4	129	14
Grade 5	93	119
Grade 6	-	14
Matter-related education		
3rd grade Inquiry Project	129	-
3rd and 4th grade Inquiry Project	93	-
No Inquiry Project	-	147

curriculum. In order to protect the students' identities, no individual demographic information was collected.

Pilot Test Scoring. Students' responses on all items were scored with two different rubrics. The first rubric was a multidimensional analytic rubric. An analytic rubric separates the evaluation criteria into distinct factors, and each factor is evaluated separately (Moskal, 2000). In this case, the "factors" were the assessment dimensions – Structure and Properties of Matter; Scale, Proportion, and Quantity; and Engaging in Argument from Evidence (see Chapter 1). The rubrics for each dimension provided a guide for matching student responses to the defined "levels" of the progress variables, and for assigning a corresponding numerical score. Each item was assigned three scores, one for each assessment dimension. To strengthen content validity, the analytic rubrics were independently reviewed by a curriculum expert before they were implemented. An example multidimensional scoring rubric is found in Figure 3.3.

The second scoring rubric was a holistic rubric, which categorized student performance based on the overall science ability demonstrated by the response. Under the holistic scoring rubric, all components of a multidimensional response were scored together as a single aggregate. The holistic rubric contained a single set of descriptors, which integrate the specific criteria from the multidimensional analytic rubrics. Under the holistic rubric, a high score was assigned to student responses that demonstrate accurate



4

Ana's block of clay

Ana has a block of clay. The block of clay is marked so that it can be divided into smaller pieces. Each smaller piece is 1 cubic centimeter.

1 cubic centimeter

1) What is the volume of the block of clay?

15 cubic centimeters

Volume:
The amount of space that something takes up

Ana takes the block of clay and molds it into a ball. She is careful not to get any air inside of the ball of clay.

2a) What is the volume of the ball of clay?

15 cubic centimeters

2b) Why do you think so? Make an argument. Give your evidence and reasoning.

the second question has the same amount of blocks as the 1st one. the first one has 15 cubic centimeters so the bottom has the same amount its just in a diffrent shape.

A

01

Figure 3.3. An example student response and scoring based on the multidimensional/analytic rubric.

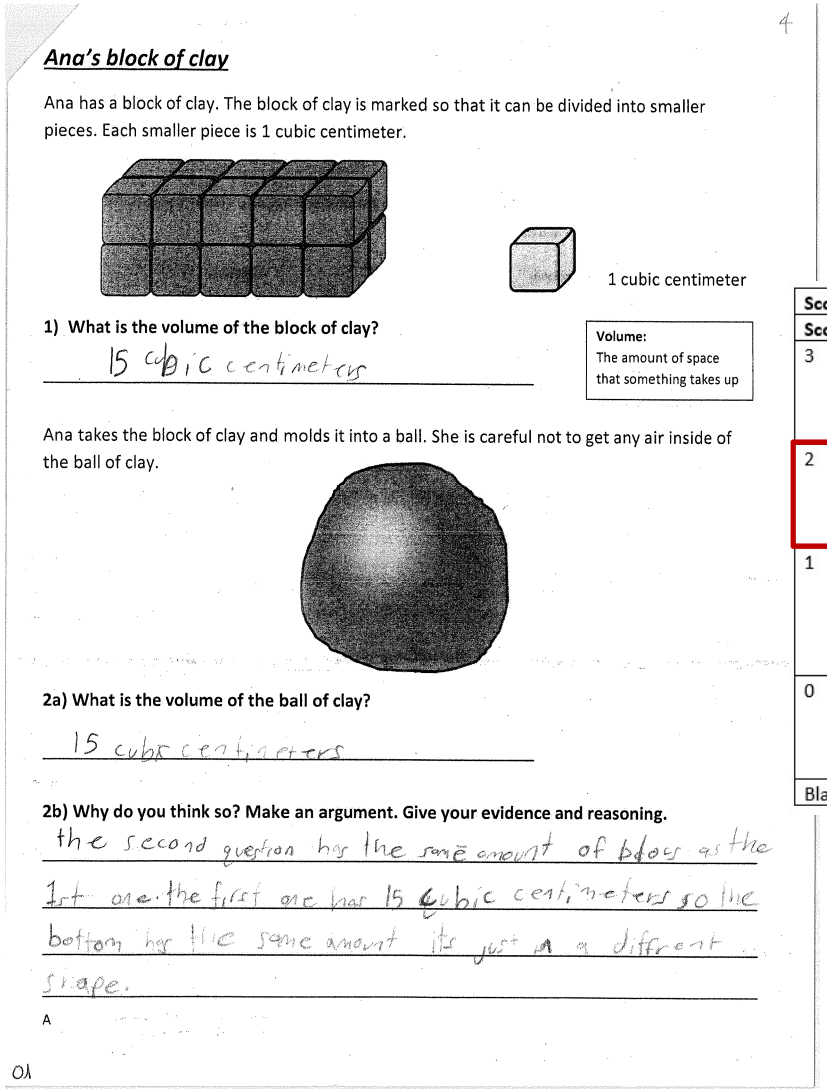
Question 1 (Scale, Proportion, and Quantity)	
Score	**Description of Response**
1 Can use centimeter cubes to measure volume.	1) 20 cubic centimeters Accept other units (cm, square centimeters) or no unit.
0 Cannot use centimeter cubes to measure volume.	1) Any other answer, besides 20.
Missing	2a) and/or 2b) are answered, but 1) is blank.
Blank	1), 2a), and 2b) are all blank.

Question 2a (Matter)	
Score	**Description of Response**
1 Understands volume of solid objects is invariant with reshaping.	Answers to 1) and 2a) are the same. OR "The same" or any close variant.
0 Does not understand volume of solid objects is invariant with reshaping.	Answers to 1) and 2a) are different. OR Bigger, smaller, different, or any close variant.
Missing	1) and/or 2b) is answered, but 2a) is blank.
Blank	1), 2a), and 2b) are all blank.

Question 2b (Argument)	
Score	**Description of Response**
3	Student supports argument with *both* evidence (observations about the block and/or the ball, or a statement that such an observation cannot be made) *and* reasoning (a clear link between the evidence and the conclusion being made, either based on the volume of solids being invariant during shape change, or some other belief). **Evidence and reasoning do not necessarily have to support a correct answer, but they should support the chosen answer.** Examples: Correct evidence and reasoning: Because Ana took the rectangle that was 20 cm and turned it into a ball, so it's just a different shape, but same volume. The block of clay was 20 cc so the ball must also be 20 cc. Incorrect evidence and reasoning: Because the ball is taller than the block so it takes up more space. The circle of clay is probably lighter, because it has no edges and is rolled up, making the material smaller and lighter.
2	Student supports argument with *either* evidence (observations about the block and/or the ball, or a statement that such an observation cannot be made) *or* reasoning (a clear link between the evidence and the conclusion being made, either based on the volume of solids being invariant during shape change, or some other statement of belief). **Evidence or reasoning do not necessarily have to support a correct answer, but they should support the chosen answer.** Examples: Correct evidence: Because it was 20 cc before she crumpled it. Incorrect evidence: The ball is squashed. I think so because about 11 full cubes could fit in the ball. Correct reasoning: Shape doesn't matter. Incorrect reasoning: Taller things take up more space.
1	Student supports a claim with irrelevant evidence (evidence that does not support their previous responses), weak evidence (appeal to authority, personal experience, tautological reasoning, or vague references to data), and/or reasoning that doesn't support their answer. **Evidence or reasoning do not necessarily have to support a correct answer, but they should support the chosen answer.** Examples: Weak evidence: Because I did it in class. Because I measured it.
0	Statements that do not offer any evidence or reasoning, e.g., "I don't know" or "It just does".
Missing	1a) and/or 2a) are answered, but 2b) is blank.
Blank	1a), 2a), and 2b) are all blank.

scientific ideas, sound measurement principles, and valid arguments. A low score was assigned to student responses that demonstrate inaccurate scientific ideas, poor understanding of measurement, and invalid or nonexistent arguments. Mid-range scores were assigned to student responses that demonstrate some combination of high and low-quality responses, with regard to the three assessment dimensions. Thus, the holistic rubrics account for response characteristics associated with the three underlying dimensions, but do not differentiate between the dimensions in scoring. Two curriculum experts offered input on scoring categories for the holistic rubric. An example holistic scoring rubric is found in Figure 3.4.

Seven raters participated in scoring student responses to the pilot test items. To gain familiarity with the scoring rubrics, all raters participated in a training exercise in which they jointly examined several written responses until they reached a consensus about the appropriate score(s) for each response. Next, raters separately examined another set of student responses, continuing until they reached agreement on 2 or more consecutive responses. The raters were then released to score pilot test responses independently. All responses were scored by at least one rater, and approximately half of responses were scored by an additional rater to facilitate an examination of interrater reliability. All student responses were scored with the multidimensional rubric first, and then the holistic rubric. Scores for all rubrics, students, items (including all item variants), and raters were collected in a single dataset.



Ana's block of clay

Ana has a block of clay. The block of clay is marked so that it can be divided into smaller pieces. Each smaller piece is 1 cubic centimeter.

1 cubic centimeter

1) What is the volume of the block of clay?

15 cubic centimeters

Volume:
The amount of space that something takes up

Ana takes the block of clay and molds it into a ball. She is careful not to get any air inside of the ball of clay.

2a) What is the volume of the ball of clay?

15 cubic centimeters

2b) Why do you think so? Make an argument. Give your evidence and reasoning.

the second question has the same amount of blocks as the 1st one. the first one has 15 cubic centimeters so the bottom has the same amount its just a diffrent shape.

Scoring rubric	
Score	**Description of Response**
3	*Exemplary* Student's response demonstrates a complete understanding of the matter concept assessed (the volume of the clay is the same after reshaping) and measurement concept assessed (calculates the volume of the clay block as 20 cubic centimeters), with clear supporting reasoning.
2	*Developing* Student's response demonstrates a partial, but incomplete understanding of the matter concept assessed (the volume of the clay is the same after reshaping) and measurement concept assessed (calculates the volume of the clay block as 20 cubic centimeters), with clear supporting reasoning.
1	*Problematic* Student's response demonstrates a partial, but incomplete understanding of the matter concept assessed (the volume of the clay is the same after reshaping) and the measurement concept assessed (calculates the volume of the clay block as 20 cubic centimeters). Supporting reasoning is unclear, weak, or nonexistent.
0	*Weak* Student's response demonstrates a poor understanding of both the matter concept assessed (the volume of the clay is the same after reshaping) and the measurement concept assessed (calculates the volume of the clay block as 20 cubic centimeters).
Blank	The student does not provide a response to any part of the question.

Figure 3.4. An example student response and scoring based on the holistic rubric.

Analysis Methods

Research question 1: To what extent does multidimensional scaffolding affect the quality of information gained from students' responses to multidimensional assessment items? Research question 1 examines the viability of items with varying amounts of multidimensional scaffolding for gathering information related to multiple dimensions of science learning. Analysis of the research question prioritizes the following evaluative criteria as crucial pieces of evidence in support of multidimensional scaffolding:

a. To what extent do students understand the intended task?
b. To what extent do student responses address all of the intended dimensions?
c. How does response time vary across items with different amounts of multidimensional scaffolding, if at all?
d. How well can the items be reliably scored?
e. How do items with different amounts of multidimensional scaffolding vary in difficulty (as estimated using the Rasch model)?
f. How well do items with different amounts of multidimensional scaffolding fit the Rasch model?

Each of these criteria is addressed separately, below.

Student understanding of the task. In order for assessment items to elicit information on multiple dimensions of science learning, it is critical that students are first able to understand the tasks required of them, especially when the tasks are complex and multifaceted (i.e., single-prompt and multiple-prompt explicit multidimensional items).

Therefore, the interview recordings were transcribed and perused for evidence of task understanding. Student comments indicative of misunderstanding or confusion about the assessment task were flagged and recorded. The different issues were grouped together by similarity and summarized, resulting in emergent qualitative categories describing 11 different sources of student confusion. This sort of descriptive coding procedure has been used previously to analyze cognitive interview data from other assessments (e.g., Howell, et al., 2013). After coding was completed, the number of students in each category was summarized and compared across individual items and aggregated to look at the distribution of task understanding across groups of items with differing amounts of multidimensional scaffolding. Although the sample size for each item is small (N as low as 4, in some cases), this information was cautiously interpreted as evidence of the relationship between the amount of multidimensional scaffolding and student understanding of the intended task.

Students address all relevant dimensions. In order to classify a student response as exhibiting proficiency on multiple assessment dimensions, the item must elicit student responses that provide information relevant to all dimensions. Using a qualitative coding procedure, students' written responses to the cognitive interview items were analyzed to determine whether a student's responses refer to any or all of the dimensions intended by the item prompt. For each item, the author identified whether or not each student's response provided sufficient information to accurately evaluate their proficiency on the Disciplinary Core Idea, Science and Engineering Practice, and/or Crosscutting Concept elicited by the item stimulus. In addition to providing information about how well the item functions, this exercise also informed preliminary rubric development for pilot test

scoring. The percentage of students providing responses that demonstrate some level of proficiency on each of the three dimensions were summarized for all tasks, and aggregated across groups of items with differing amounts of multidimensional scaffolding. Again, sample size for each item is small, but these summaries provide information to support inferences about the utility of multidimensional scaffolding to gather information relevant to multiple dimensions.

Pilot test data was also examined to determine how frequently students addressed each of the three dimensions for each of the different item scaffolding variations. When the multidimensional scoring rubric was used, raters were given the option to assign a score of "Missing on [Dimension]" to a particular item and dimension. For multiple-prompt items, this occurred when a student left one or more dimensional sub-prompts blank, but provided an answer to at least one of the sub-prompts within an item context. For single-prompt items, this occurred when a student provided a response that did not provide any information about their ability on one or more of the assessment dimensions. Raters were instructed to mark a response as "Blank on All Dimensions" if the student did not respond to any part of the prompt/sub-prompts. "Missing on [Dimension]" and "Blank on All Dimensions" responses were counted, and the percentage of both types of missing responses was aggregated and compared across all multiple-prompt and single-prompt items.

Response time. Time requirements are an important consideration for any assessment, as children's instructional time is valuable and lengthy assessments may fatigue or overburden them. Comparing the length of time for completion across items facilitates test construction that considers the costs and benefits associated with

lengthening the students' testing burden. Based on recorded observations from the cognitive interviews, average response time was calculated for each item and summarized across groups of items with differing amounts of multidimensional scaffolding. Due to the small sample size of students encountering each cognitive interview item, statistical comparison of response time (e.g., t-tests) are likely to lack the power necessary to observe differences between different classes of items. Therefore, average response time for different groups of items were compared holistically and cautiously interpreted.

Scoring reliability. Scoring reliability is essential for valid interpretations of assessment performance. Whenever subjective judgment is involved in scoring there is the potential for measurement error. Information about interrater reliability provides an indication of how much measurement error is introduced during the scoring process (AERA, 1999). Examining the scoring reliability of items with different levels of multidimensional scaffolding provides crucial evidence to support the choice of reliable item types in the final instrument. After pilot administration and scoring, each item's interrater reliability was assessed by calculating the intraclass correlation coefficient (ICC; Shrout & Fleiss, 1979). The specific form of the ICC used was what Shrout and Fleiss refer to as ICC(2,1) (p. 423), which is the reliability of a single rating estimated from a two-way random-effects analysis of variance. This statistic is appropriate when each observation is rated by the same *k* raters, who are assumed to be a sample from a larger population of possible raters. The ICC ranges from 0 to 1 and indicates the degree of similarity among ratings, with high values signifying that interrater reliability is good. The ICC was examined for each individual item, and items with a high proportion of unique variance (ICC<~0.7) were flagged for review.

Since the ICC is a correlation coefficient, the distribution of the sample correlation coefficients will always be non-normal (Fisher, 1915). Therefore, all ICC's were transformed to a normal distribution using Fisher's Z-transformation in order to compute the average ICC for all items within a multidimensional scaffolding or response format group.

Item difficulty. Pilot test data were also used to examine differences in item difficulty among items with differing levels of multidimensional scaffolding. Student performance data for all 64 items (20 item contexts with 3-4 dimensional sub-prompts per item context) was scaled using ConQuest 4 (Adams, Wu, & Wilson, 2015). Marginal maximum likelihood estimation was used to generate item difficulty estimates and fit statistics. Wright maps were used to visually compare the distribution of item thresholds for a) multiple-prompt explicit multidimensional and single-prompt explicit multidimensional versions of the same items, and b) selected-response and open response versions of the same items, and patterns in the relative difficulty of different item variants were observed.

Item fit. Pilot test data was also used to examine item fit across each group of items with differing levels of multidimensional scaffolding. Item fit is a measure of how well the overall pattern of item responses conforms to the requirements of the Rasch model. In particular, item fit statistics provide an indicator of the size of the item residuals (deviations between observed and expected responses). For this analysis, weighted mean square (MNSQ) statistics generated by ConQuest 4 (Adams, Wu, & Wilson, 2015) were examined as indicators of item fit. The impact of individual outliers is attenuated in the weighted mean square, so it was used in place of the unweighted

statistic. Items with consistently large residuals generate a large weighted mean square statistic, indicating that actual student responses deviated significantly from expected. Items with consistently small residuals generate a small weighted mean square statistic, indicating that actual student responses conformed to the expected responses. In general, weighted mean square values between 0.75 and 1.33 are considered acceptable. A t-statistic is also generated, which indicates how likely it is to observe a given weighted mean square statistic, based on a normal distribution. The weighted mean square and t-statistics were examined in conjunction, and items with both a) weighted mean square values falling outside the range of 0.75 to 1.33, and b) t-statistics with absolute values larger than 1.96 (i.e. mean square with less than a 5% chance of observation) were flagged. The frequency of flagged items examined across groups of items with differing levels of multidimensional scaffolding, informing a holistic evaluation of relative item fit for each group.

Research question 1a: Does the impact of scaffolding and/or response format vary for students of different abilities? Research question 1a examines the effect of assessment scaffolding and response format on student performance for students of differing abilities. Analysis of the research question examines the following evaluative criterion:

a. How do student response rates to the scaffolding variations differ for students of different abilities?
b. How do differences in interrater reliability on the item variations relate to student ability?
c. How do differences in item estimates between the item variations relate to student ability level?
d. How do differences in item fit on the item variations relate to student ability?

This criterion is addressed, below:

Difference in missing responses for students of varying abilities. The sample of students was divided into three groups based on ability – high, medium, and low – as measured by the WLE estimates generated by ConQuest. Within each of these ability groups, the percentage of responses classified as "Missing on [Dimension]" and "Blank on All Dimensions" was calculated for each assessment dimension and scaffolding variation. This was done by counting the total number of responses scored as "Missing on [Dimension]" or "Blank on All Dimensions" on a particular item and dimension, summing across all item variants of the same type (i.e., all multiple-prompt items and all single-prompt items), and dividing by the number of possible responses to all items of that type (calculated based on the number of students who saw each item according to the assignment of test forms). Then, the percentage of missing responses was compared across low, medium, and high ability students on single-prompt and multiple-prompt items. The difference between a subgroup's response rate to single-prompt and multiple-

prompt items were attributed to the difference in amount of scaffolding, and variations in this difference were examined across the different ability groups.

Differences in interrater reliability among low, medium, and high ability students. The same ability subgroups were used as those in the previous analysis. In those subgroups, only a subset of responses were double scored. Based on these responses, the intraclass correlation coefficient (ICC) was calculated for each item. ICC's were transformed via Fisher's z-transformation, and then averaged across each item variation (single-prompt, multiple-prompt, constructed-response, and selected-response). These averages were compared across item variations and subgroups.

Differences in item difficulty across low, medium, and high ability students. Each of the ability subgroups were reanalyzed with a multidimensional Rasch model. This resulted in a separate item/threshold estimate for each ability group. As in the previous analysis, differences in item difficulty among item variations with different levels of scaffolding and response formats were assessed by examining trends in item difficulty on an item map. Differences in item difficulty between variations were examined, and these differences were then compared to see whether they held constant across ability groups.

Differences in item fit across low, medium, and high ability students. Based on the results of the subgroup Rasch analysis, item fit was examined again to determine whether any patterns in misfit were related to student ability. Average fit statistics

(weighted mean square and T-statistics) were computed for each item variation and subgroup.

Research question 2: To what extent do item response formats affect the quality of information gained from students' responses to multidimensional assessment items? Research question 2 deals with the efficacy of selected-response and open-response item formats for gathering information related to multiple dimensions of science learning. Analysis of the research question prioritizes the following evaluative criteria as crucial pieces of evidence in support of each item response format:

a. To what extent do students understand the intended task?
b. How does response time vary across the response formats?
c. How well can the items be reliably scored?
d. How do selected-response and constructed-response items differ in difficulty (as estimated using the Rasch model)?
e. How well do selected-response and constructed-response items fit the Rasch model?

Each of these criteria is addressed separately, below.

Student understanding of the task. Student understanding of the task was measured during the second round of cognitive interviews, this time with a focus on uncovering differences in task understanding that arose when a selected-response format was utilized. Qualitative codes were applied to the interview transcripts and subsequently categorized and summarized, as in the analysis for Research Question 1, except with a focus on examining differences across response formats.

Response time. Response time was also used to provide information about the temporal burden associated with each response format. As in the analysis for research question 1, average response time was calculated for each item and summarized across each response format.

Scoring reliability. After the pilot test was administered, student responses were scored. All responses were scored by raters, including responses to selected-response items. The intraclass correlation coefficient (ICC) was calculated as a measure of scoring reliability for all assessment items. The ICC was averaged across all selected-response and open response items via Fisher's Z-transformation, and scoring reliability was compared. As the scoring reliability of an open-ended item decreases, selected-response alternatives become more attractive options to maintain an instrument's reliability.

Item difficulty. Pilot test data was also used to examine differences in item difficulty among selected-response and open-response item formats. Identical to the item difficulty analysis for Research Question 1, item difficulty estimates were compared by visually examining the placement of matching thresholds from selected-response and open-response versions of an item on a Wright map, and looking for patterns in item difficulty among the two variants. As above, item difficulty was also examined across subgroups of students with low, medium, and high ability.

Item fit. Pilot test data was also used to examine item fit across each of the item response formats. As in the analysis for Research Question 1, weighted mean square and the associated t-statistic from the model output provided by ConQuest 4 (Adams, Wu, & Wilson, 2015) were examined as indicators of item fit. As above, fit was examined again across subgroups of students with low, medium, and high ability.

Research question 3: To what extent do unidimensional and multidimensional scoring and modelling approaches affect the empirical relationships among the 3 dimensions of science learning (assuming that such relationships exist)? Research question 3 deals with selecting the appropriate scoring method and psychometric model to produce estimates of student performance. The Next Generation Science Standards (NGSS Lead States, 2013), on which instrument development was based, have an explicit three-dimensional structure. The construct definition therefore implies that student performance should be scored and scaled multidimensionally. This assumption was tested by comparing three different scoring rubric/ measurement model combinations: 1) a multidimensional scoring rubric combined with a multidimensional measurement model; 2) a multidimensional scoring rubric combined with a unidimensional measurement model; and 3) a holistic (unidimensional) scoring rubric combined with a unidimensional measurement model. The resulting model fit statistics provide information about the underlying structure of an NGSS-based elementary science assessment, and consequently, the most appropriate way to score and interpret student responses.

Prior to examining the dimensionality of the multidimensional and holistic datasets, several different model variations were compared in order to find the most parsimonious baseline model that fit all item parameters, step parameters, and rater effects. Each of the three dimensional subtests was initially analyzed as its own scale. Different raters scored each item, meaning that comparisons between items that utilized different scaffolding and response formats may have been compounded by rater effects. In particular, the Argument items were affected by low interrater reliability. In order to

make comparisons among the different scaffolding and response format variations, a multifaceted model was employed to account for differences in raters' application of the Argumentation rubric. The Matter and Scale, Proportion, and Quantity dimensions tended to have much higher interrater reliability; therefore, a multifaceted approach was not necessary on these subtests. ConQuest 4 software (Adams, Wu, & Wilson, 2015) was used to estimate all models. ConQuest uses the MRCMLM (Adams, Wilson, & Wang, 1997) to concurrently scale both dichotomous and polytomous items on one or more assessment dimensions. ConQuest was also used to model rater effects using multifaceted models.

To evaluate dimensionality, scores from the multidimensional scoring rubric were scaled twice: (1) a unidimensional scale representing students' overall science learning, (2) a between-item multidimensional analysis of three NGSS-aligned subscales as related dimensions, in which student performance on one dimension contributes information to the estimation of their ability on the other dimensions. Scores from the holistic scoring rubric were scaled once under a unidimensional model. Person parameter estimates were subsequently generated using both weighted maximum likelihood (WLE) (Warm, 1989) and Bayesian expected a posteriori (EAP) estimation procedures. In multidimensional IRT models, WLE's have been shown to display less bias than maximum likelihood person ability estimates (MLE's) and EAP estimates, although EAP estimates are also associated with smaller standard errors (Wang, 2015). Therefore, both WLE's and EAP's were examined at various stages in the analysis.

For each model, ConQuest computed parameter estimates, standard errors, and fit statistics for both item difficulty and person ability, as well as dimensional means,

variances, and covariances (Adams, Wu, & Wilson, 2015). All of these indicators were used to provide evidence supporting the underlying structure of the instrument.

Although the main purpose of Research Question 3 is to examine the underlying structure of the assessment and its relationship with scoring decisions, the analysis takes into account several indicators of technical quality of the item and person estimates, in addition to a dimensional analysis of the overall instrument. This is because the quality of the items and the interpretations derived from student responses are a critical component of any further argument about the instrument's internal structure (Chapelle, Enright, & Jamieson, 2010; Peoples, 2012). Structural validity evidence about the instrument as a whole is irrelevant if the function of individual items and/or interpretability of person estimates is poor. Thus, analysis of Research Question 3 prioritizes the following evaluative criteria as crucial pieces of evidence about the instrument's structural validity:

a. Which model demonstrates best fit, according to the deviance statistics reported by ConQuest 4 software?
b. How well do the items fit the scale(s) under each model?
c. What are the correlations between ability estimates from the NGSS's 3 dimensions of science learning?
d. How well do student ability estimates fit each model?
e. How do the covariances between the NGSS dimensions of science learning vary for students of different ability levels?
f. How do the reliability estimates vary for each scale and subscale?
g. How precise are item difficulty estimates under each model?
h. How precise are person ability estimates under each model?

Model fit. Using the multidimensional scoring rubric, the unidimensional model is nested within the multidimensional model. This means that the difference in each model's deviance statistic (G^2) can be referred to a Chi-squared distribution with the appropriate degrees of freedom (Adams, Wilson, & Wang, 1997). The degrees of freedom used to calculate the Chi-squared critical value is based on the difference in the number of parameters in each model, and a statistic larger than the critical value indicates a significant difference in model fit. A smaller deviance statistic indicates better model-data fit; therefore, if the difference between two models' deviance statistics is statistically significant, we can conclude that the model with the smaller deviance statistic fits the data better. Overall model fit is an indicator of the degree to which the assumptions of the hypothesized psychometric model hold up for a particular set of test data, and thus provides information about the overall appropriateness of a unidimensional or multidimensional interpretation of student responses.

Note that model fit can only be directly compared for models based on the same dataset. Thus, the fitted model based on data from the holistic (unidimensional) scoring rubric could not be directly compared to either of the other models by comparing the difference in model deviances (G^2) to the Chi-squared distribution.

An overall assessment of model fit was based on triangulation of the evidence provided by all of the previous analyses.

Item fit. Item fit was examined across the three hypothesized models. As in the analyses for Research Questions 2 and 3, weighted mean square fit statistics and their associated t-statistics were examined as indicators of item fit. Whereas the previous analyses focused on comparing item fit across different groups of items with different

features, this analysis focuses on the overall fit of all items within each of the hypothesized models.

Correlation between NGSS dimensions. Correlations between latent dimensions are estimated directly during estimation of the multidimensional model (Adams, Wilson, & Wang, 1997). Multidimensional correlations are more accurate than dimensional correlations from a joint unidimensional analysis, because the joint unidimensional correlations are attenuated by measurement error (Adams, Wilson, & Wang, 1997; Briggs & Wilson, 2003). A moderate correlation between dimensions improves the precision of the person and item estimates; however, a high correlation implies that the dimensions may not be distinct after all (Adams, Wilson, & Wang, 1997). Therefore, correlations between the three latent dimensions of the NGSS, as reported by ConQuest 4, were examined. The sizes of the correlations serve as evidence of the appropriateness of a unidimensional or multidimensional interpretation of assessment data. Furthermore, correlations were also examined among subgroups of low, medium, and high ability students to see whether the strength of the relationships between dimensions varied depending on student ability.

Person fit. Similar to item fit, person fit statistics were also reported for each case estimate (or in the case of the multidimensional model, set of case estimates). Person fit was examined across the three hypothesized models. Mean square statistics were examined, and cases with values less than 0.7 or more than 1.3 were flagged as misfitting (Wright & Linacre, 1994), and the numbers of misfitting cases were compared across all three models.

Heteroscedasticity/homoscedasticity of covariances between dimensions. One more analysis was conducted to explore the nature of the relationship between the model dimensions; specifically, whether it is homoscedastic or heteroscedastic. Previously, student ability estimates from multidimensional assessments have been shown to have heteroscedastic interactions (Brown, Castle, & Chappe, 2014). Heteroscedasticity implies that the relationship among dimensions differs at different levels of each construct; for example, that the relationship among the dimensions is stronger among students with lower ability estimates on the three dimensions and weaker among students with higher ability estimates. Heteroscedasticity indicates a potential discrepancy between the intended construct and the actual construct being measured, and thus may affect the validity of interpretation of student estimates. To examine this issue, two-dimensional scatterplots were generated to plot estimated student abilities with respect to pairs of dimensions. These scatterplots were examined for evidence of heteroscedasticity, which is often indicated by differences in the amount of variation at different points along the trend line.

Reliability. Reliability statistics provide a concise summary of the amount of error in the estimates produced by a test (Wilson, 2005). Theoretically, reliability measures how consistently the instrument classifies examinees (Russell & Airasian, 2012). Instruments with high reliability consistently provide a similar measure or score for an individual every time they encounter the instrument.

The separation reliability coefficient is a measure of internal consistency that was used to evaluate scale reliability. The separation reliability coefficient is

$$r = \frac{Var(\theta)}{Var(\hat{\theta})},$$

where $Var(\theta)$ is the variation in examinees' locations explained by the model, and $Var(\hat{\theta})$ is the total observed variation in examinees' estimated locations (Wilson, 2005). The difference between $Var(\theta)$ and $Var(\hat{\theta})$ is the amount of variance attributable to error. High values of the separation reliability coefficient (greater than ~0.80) indicate that the amount of error variance in student ability estimates is low.

Reliability tends to decrease as the number of items contributing to a scale decreases; therefore, it is usual for the reliability of multidimensional subscales to be lower than the reliability of a unidimensional scale (Briggs & Wilson, 2003). ConQuest produces separation reliability estimates computed from both the WLE and EAP person ability estimates. Although WLE estimates are less biased than Bayesian estimates (Warm, 1989), the EAP estimates tend to be more precise estimates of student ability (Wang, 2015) and thus will likely have a higher reliability. The WLE and EAP separation reliability coefficients were compared across both models to investigate the decrement in reliability associated with separating science learning into three NGSS-aligned subscales. The size of the difference between WLE and EAP reliability was also taken into account when making recommendations on the most appropriate person ability statistic for the final model. Finally, scale reliability was also evaluated across subgroups of low, medium, and high ability students to see whether there was any relationship between student ability and scale reliability.

Precision of item difficulty estimates. An item difficulty statistic is only an estimate, meaning that there is some uncertainty about the true item location. The degree of uncertainty is captured in the standard error of measurement (Wilson, 2005). The size of the standard error is an indicator of how accurately the item is able to place examinees

on the scale. In fact, the inverse of the standard error is a statistic referred to as the item information function, signifying that it is an indicator of the amount of information that an item contributes to an examinee's score (de Ayala, 2009). In this analysis, the standard errors of the item difficulty estimates were examined across each of the modeling approaches in order to evaluate the extent to which the scoring rubric and model dimensionality affect the standard errors.

Precision of person ability estimates. Standard errors of the person ability estimates were examined across all models. In the multidimensional model, the subscales have fewer items than when all items are placed on a unidimensional scale. Tests with smaller numbers of items tend to have larger errors (Briggs & Wilson, 2003). Although information about the relationship between the dimensions should partially mitigate this phenomenon, there remains some concern about the amount of error in the multidimensional person estimates compared to that in the unidimensional person estimates. Therefore, WLE and EAP ability estimates from all models were examined in order to investigate the extent of the increase in error of the multidimensional estimates relative to the unidimensional estimates.

Research question 4: How well does the assessment function in its intended purpose of measuring student proficiency on the construct(s) defined by the *Inquiry Project* curriculum? Research question 4 addresses some remaining validity considerations, including the alignment between item estimates and the three hypothesized underlying constructs, the functioning of individual items as measured by fit statistics, alignment between items and student ability, DIF, and differences in student

performance based on grade level and Inquiry Project participation. The following criteria provide information about these aspects of assessment validity:

a. What is the range of item difficulty estimates under each model?
b. What is the range of person ability estimates under each model?
c. To what extent does the order of item difficulty estimates align with the hypothesized ordering of levels in the construct map?
d. How well do assessment items fit the item response model?
e. How precise are student ability estimates on each dimension?
f. Does item difficulty vary for students in different subgroups? Specifically, are there differences in item difficulties depending on grade level or whether students participated in the Inquiry Project curriculum?
g. How are *Inquiry Project* participation and grade level related to student performance?

Each of these criteria is addressed separately, below.

Range of item difficulty estimates. Item difficulty estimates and person ability estimates were evaluated concurrently for each model with a Wright map. The range of item difficulty estimates indicates whether an assessment provides reliable estimates of ability at all levels of the construct. Items are most informative for an examinee when their locations are in the neighborhood of the examinee's ability (Wilson, 2005). Therefore, to adequately measure the full range of examinees there should be a number of items located in each region of the scale. A wide range of item difficulty estimates offers additional benefits for measurement, including the ability to capture growth over time. Furthermore, the visual map of item difficulty estimates can be compared with the

predicted order of item difficulty based on the theory set forth in the construct maps, allowing researchers to verify the theorized construct (Wilson, 2005).

This analysis focuses in particular on the effect of separating the dimensions when a multidimensional scoring rubric is used. One obvious consequence of defining a measurement model with different subscales is that the number of items on each scale is only a fraction of the total number of items. A substantial decrease in number of items may lead to exposure of gaps on the individual subscales. Ideally, the range and distribution of item estimates on each subscale is preserved when the dimensions are separated; however, this was not anticipated. Evaluation of item difficulty compared the range of item difficulty estimates under the unidimensional model with the ranges of item difficulty under the multidimensional model, to evaluate whether the latter model lost sensitivity at the extremes of the measurement scales.

Range of person ability estimates. Weighted likelihood estimates (WLE's; Warm, 1989) were used to assess the range of person abilities. Person ability estimates were examined alongside the item difficulty estimates using a Wright map. It is important to examine the range of person location estimates because the amount of variation in person ability estimates has implications for both a) the items' functionality, and b) the structure of the underlying construct. A substantial range of person estimates indicates both that the items are successfully differentiating among students of differing abilities and that examinees demonstrate variability on the underlying construct. (A small range of person estimates may indicate a failure in either or both the items design and the definition of the construct.) In the case of subscale measurement, a substantial range of person ability estimates indicates that the subscales are worth measuring individually, as

opposed to being collapsed into an overall science scale. Therefore, the ranges of person ability estimates were compared across modeling approaches.

Order of item difficulty estimates. Each of the three sub-constructs has been operationalized as a progress variable, or a construct map similar to a learning progression. The construct map is a hypothesis; it is a best guess based on previous research and knowledge about how children progress in developing the knowledge and skills that fall under the domain of the construct. One way to evaluate the validity of the hypothesized construct is to determine whether the results of the assessment match the anticipated results based on the hypothesis. Since each assessment item was written to differentiate between two or more levels of each progress variable, the anticipated order of item difficulties is easy to infer from the progress variables (see Chapter 1). For each dimension, the Wright map was examined, anticipated and actual orders were compared, and discrepancies were noted and interpreted in terms of the definition of the underlying construct.

Item misfit. Item misfit was also examined as part of Research Questions 1 and 2, to examine whether there might be patterns in item fit that reflect item characteristics like scaffolding and/or response format. Here, it will be examined again, this time as an indicator of the functionality of individual items. Item fit, measured by weighted mean squares and their associated t-statistics, indicates the extent to which student responses conform to what is expected, based on item difficulty and person ability estimates from the item response model. When students perform unpredictably, items have large fit statistics (MNSQ > 1.33, T > 1.96), and when student responses are highly predictable, items have small fit statistics (MNSQ < 0.75, T < -1.96). In each case, the pattern of

responses may be caused by item-specific features that influence the way students interact with the item. Any flagged items were closely examined to look for item features that may unintentionally interfere with measurement of the intended construct.

Precision of person ability estimates. Like the item difficulty estimates, person ability estimates also have associated standard errors of measurement. The standard error captures the uncertainty of the estimate, and its size tells us about the amount of uncertainty in each estimate. Examinees with ability estimates near the population mean (i.e., nearer the location of a majority of items) tend to have smaller standard errors than examinees at the ends of the ability range (Wilson, 2005). The standard error is a useful indicator of test quality because it provides information about how well the assessment is able to pinpoint a student's particular level of ability relative to other students. Standard errors of student ability estimates were examined to see whether there any areas (i.e., a particular dimension or region of a dimensional subscale) where precision could be improved.

Differential item functioning (DIF). DIF refers to the existence of one or more factors, besides person ability and item difficulty, which contribute systematically to the probability of examinee responses. In this analysis, two grouping variables – grade level and completion of the Inquiry Project curriculum – were examined for evidence of DIF.

ConQuest 4 (Adams, Wu, & Wilson, 2015) examines DIF by adding the grouping variables as additional parameters in the item response model (Adams & Wu, 2010). For dichotomous grouping variables, the resulting model contains estimates for each item separately for each subgroup. Then the difference between estimates can be compared and evaluated for significance, based on the standard error of the estimates. If there is a

significant difference between the subgroup estimates for an item, this is evidence of DIF. An overall group mean is also reported, which is separate from the DIF analysis, but demonstrates whether there is a significant difference between each subgroup's performance on the overall assessment.

For polytomous grouping variables, two alternative item response models are generated – one that accounts for the subgroup identity, and one that does not account for subgroup identity. If the model that accounts for subgroup identity has a significantly better overall fit to the data than the model that does not account for subgroup identity (as measured by the difference in model deviance statistics referred to a Chi-square distribution), then this is interpreted as evidence of DIF.

Inquiry Project *participation and grade level.* Box plots were created to visually compare the variance in student ability estimates within classrooms that participated in the pilot test. This analysis focused specifically on looking for differences in within-classroom variance among the following subgroups: a) Inquiry Project and non-Inquiry Project classrooms, and b) 4^{th} and 5^{th} grade classrooms. The average variance was computed for each of these 4 subgroups and 3 assessment dimensions, with class size used to weight the influence of each variance estimate on the average

In this analysis, students in the same classroom share similarities, meaning that classroom within-group variance may be underestimated. To account for this possibility, an unconditional multilevel model was employed to generate corrected estimates of within-group variance for each subgroup. These corrected estimates were compared via an F-test, using the corrected degrees of freedom provided by HLM software (Scientific Software International, 2017). The ICC, which indicates the proportion of variance in the

outcome that is attributable to classroom membership, was also examined for each subgroup and outcome variable.

This analysis provides information about whether *Inquiry Project* curriculum participation and grade level affect variability in student performance, while taking into account the variance shared between students in the same classrooms.

Chapter 4: Results

Results of the data analysis procedure outlined in Chapter 3 are presented here, organized by research question and sub-question. Research questions 1 and 2 utilized many of the same criteria to examine multidimensional scaffolding and response format, and therefore results are reported together.

Research questions 1 & 2: To what extent do multidimensional scaffolding and response format affect the quality of information gained from students' responses to multidimensional assessment items?

Student understanding of the task. Student understanding of the task was investigated through the analysis of transcripts from both rounds of cognitive interviews. Each cognitive interview was recorded and transcribed. Transcripts were reviewed, and any sources of student confusion were identified and categorized. A student's response to a single item could receive multiple codes if they expressed multiple sources of confusion.

Multidimensional scaffolding. Three different types of items were created by varying the amount of scaffolding. These three variations comprised: multiple-prompt explicit multidimensional items, single-prompt explicit multidimensional items, and single-prompt implicit multidimensional items[17]. Ten different item contexts, or scenarios, were utilized, and three item variations were created for each of these ten contexts, resulting in a starting set of 30 items. An example of the three item variations can be found in Figure 3.1.

[17] See Chapter 3 for a detailed description of the multidimensional scaffolding variations.

These 30 items were tested with 26 students during Round 1 of the cognitive interview process, such that 4-5 students saw each item. After reviewing the transcripts from all of the Round 1 cognitive interviews, 11 sources of misunderstanding were identified (Figure 4.1).

The distribution of the 11 sources of misunderstanding was examined across the 3 types of items for each item context individually and in the aggregate. A table describing the frequency of all 11 issues within each item context can be found in Appendix B. At the item context level, the presence or absence of certain issues varied widely. For example, "unfamiliarity with task context" was only observed for one of the item contexts: an item which asked students to compare the weights of an equal amount of coffee beans and ground coffee. Not all students were familiar with coffee beverages. Students appeared to be comfortable with the task context of all other items. On the other hand, "misunderstanding or misinterpretation of critical piece of task" was present for 7 out of 10 item contexts, suggesting that this was a much more common issue for students.

Sources of Misunderstanding – Multidimensional Scaffolding

1. Misunderstanding or misinterpretation of critical piece of task
2. Overlooks critical piece of information
3. Unfamiliarity with task context
4. Unfamiliarity with item vocabulary (non-scientific)
5. Unfamiliarity with measurement unit
6. Unfamiliarity with measurement tool
7. Unfamiliarity with measurement calculation
8. Misunderstanding of key concepts
9. Misunderstanding of visuals
10. Confused by item layout
11. Provides correct answer despite avoiding intended task

Figure 4.1. Eleven observed sources of student confusion during cognitive interviews with multidimensional scaffolding item variations.

Several of the 11 sources of misunderstanding were unlikely to be influenced by the amount of scaffolding present in the items due to the nature of the issue; for instance, a misunderstanding of a key concept is caused by the student's proficiency with an item's underlying topic and is not likely to be affected by the amount of item scaffolding. Scaffolding affects the rate at which information about the task is presented to the student, and certain issues are more likely to be affected by this than others. An issue like "overlooks a critical piece of information" may be exacerbated when students are presented with a large amount of information about the task at once, as they are in the single-prompt multidimensional version of the item. This analysis focuses mainly on only three of the issues of misunderstanding, which are the most likely to be affected by the amount of scaffolding present in the item: "overlooks a critical piece of information"; "confused by item layout"; and "misunderstanding or misinterpretation of critical piece of task."

The frequencies of these 3 sources of misunderstanding across all task contexts can be found in Table 4.1. For example, the first row describes the number of observations of "Misunderstanding or misinterpretation of critical piece of task" across the three scaffolding variations for all items: 23% of students misunderstood or misinterpreted some critical aspect of the task when a multiple-prompt version of the item was used, 26% of students did so with a single-prompt multidimensional version of the item, and 25% did so with a unidimensional version of the item. This category was assigned when a student indicated confusion about the nature of a task, or performed a different task than intended. For example, on the item "Carol's butter" (see Appendix F), the intended task asked students to determine the weight of a sample of butter after it had

Table 4.1.

Frequency of Sources of Misunderstanding Affected by Scaffolding

	Multiple-prompt multidimensional		Single-prompt multidimensional		Unidimensional	
Issue	Count	Percentage	Count	Percentage	Count	Percentage
Misunderstanding or misinterpretation of critical piece of task	10	23.26%	11	25.58%	11	25.00%
Overlooks a critical piece of information	1	2.33%	5	11.63%	1	2.27%
Confused by item layout	1	2.33%	2	4.65%	3	6.82%
Total number of items viewed by students during Round 1*	43	100.00%	43	100.00%	44	100.00%

*Indicates the total number of times that any student encountered an item in each multidimensional scaffolding category. For example, in the multiple-prompt category, 26 students each saw at least one multiple-prompt item, and some saw more than one. This led to a total of 43 instances of students encountering multiple-prompt items.

been melted. However, many students indicated confusion about whether or not they were supposed to account for the weight of the container holding the melted butter, which interfered with their ability to demonstrate their reasoning about the weight measurement.

Although there are differences in the frequency of certain issues for different levels of multidimensional scaffolding, these differences are hard to interpret because of the small number of students who responded to each item. Therefore, differences between items with different amounts of multidimensional scaffolding may be attributable to differences among the students who saw each item, and should be interpreted cautiously.

Overall, there was not a difference in the frequency of student misunderstanding or misinterpretation of critical pieces of the task among the different levels of multidimensional scaffolding. Students were more likely to overlook a critical piece of information when they encountered a single-prompt multidimensional item (12% of the

time) than they were for either of the other two scaffolding levels. Students were not likely to be confused by item layout in any of the three scaffolding categories.

Response format. Three new types of items were created by varying both scaffolding and response format. These three variations included: constructed-response multiple-prompt items (same as the multiple-prompt multidimensional item from Round 1, but revised), selected-response multiple-prompt explicit multidimensional items, and selected-response single-prompt implicit multidimensional items. An example of the two new item variations can be found in Figure 4.2. The same ten item contexts were utilized. Two of the item contexts did not include a constructed-response multiple-prompt item variation in Round 2, as there were no changes to the items after Round 1 and therefore no reason to seek input from additional students. This left 28 items, which were then tested with 37 students during Round 2 of the cognitive interview process such that 5-8 students saw each item.

After reviewing the transcripts from all of the Round 2 cognitive interviews for evidence of student understanding of the tasks, 16 sources of misunderstanding were identified. These included 10 of the same sources from the Round 1 cognitive interviews and 6 additional sources of misunderstanding (Figure 4.3).

The distribution of the 16 sources of misunderstanding was examined across the 3 types of items for each item context individually and in the aggregate. A table describing the frequency of all 16 issues within each item context can be found in Appendix B. At the item context level, the presence or absence of certain issues varied widely. For example, "large/small scale makes problem difficult to think about" was observed in only two of the item contexts which required mathematical reasoning with small fractions. On

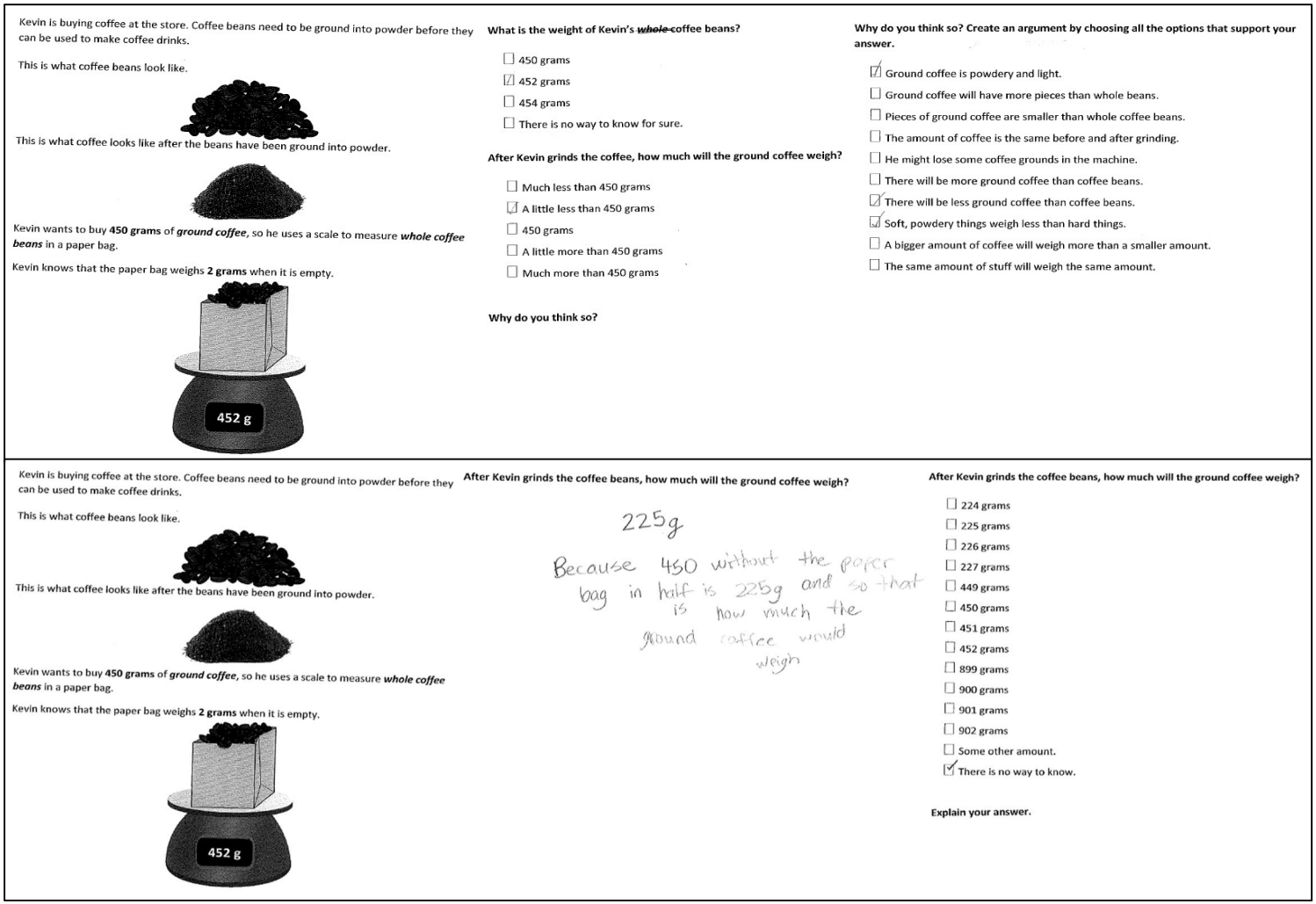


Figure 4.2. Selected-response item variations. The top panel shows a selected-response multiple-prompt explicit multidimensional item, and the bottom panel shows a selected-response single-prompt implicit multidimensional item.

Sources of Misunderstanding – Response Format

1. Misunderstanding or misinterpretation of critical piece of task
2. Overlooks critical piece of information
3. Unfamiliarity with task context
4. Unfamiliarity with item vocabulary (non-scientific)
5. Unfamiliarity with measurement unit
6. Unfamiliarity with measurement calculation
7. Misunderstanding of key concepts
8. Misunderstanding of visuals
9. Confused by item layout
10. Provides correct answer despite avoiding intended task
11. Answer options don't reflect student understanding
12. Overwhelmed by amount of information/answer choices
13. Alternative explanation based on extraneous factor
14. Answer options lead student to rethink answer
15. Large/small scale makes problem difficult to think about
16. Misunderstanding of answer options

Figure 4.3. Sixteen observed sources of student confusion during cognitive interviews with response format item variations.

the other hand, "misunderstanding or misinterpretation of critical piece of task" was present for all 10 of the item contexts, suggesting that this was a much more common issue for students.

As in Round 1, several of the 16 sources of misunderstanding were unlikely to be influenced by either scaffolding or response format due to the nature of the issue. This analysis focuses mainly on five issues that were likely to be affected by response format: "answer options don't reflect student understanding"; "overwhelmed by amount of information and/or answer choices"; provide 'correct' answer despite avoiding intended task"; "answer options lead student to rethink answer"; and "misunderstanding of answer options." In addition, the three sources of misunderstanding identified as likely to be affected by scaffolding from Round 1 were reexamined with the Round 2 data

("overlooks a critical piece of information"; "confused by item layout"; and "misunderstanding or misinterpretation of critical piece of task").

The aggregated frequencies of the 5 sources of misunderstanding likely to be affected by response format can be found in Table 4.2. Again, differences in the frequency of certain issues for different response formats are hard to interpret because of the small number of students who responded to each item. Therefore, differences between items with different response formats may be attributable to differences among the students who saw each item, and should be interpreted cautiously.

Table 4.2.

Frequency of Sources of Misunderstanding Affected by Response Format

	Constructed response/Multiple prompt explicit multidimensional		Selected response/Multiple prompt explicit multidimensional		Selected response/Single prompt implicit multidimensional	
Issue	Count	Percentage	Count	Percentage	Count	Percentage
Answer options don't reflect student understanding	0	0.00%	8	11.43%	13	22.03%
Overwhelmed by amount of information and/or answer choices	1	2.13%	6	8.57%	4	6.78%
Provide "correct" answer despite avoiding intended task	9	19.15%	11	15.71%	10	16.95%
Answer options lead student to rethink answer	0	0.00%	9	12.86%	9	15.25%
Misunderstanding of answer options	0	0.00%	2	2.86%	1	1.69%
Total number of items viewed by students during Round 2*	47	100.00%	70	100.00%	59	100.00%

*Indicates the total number of times that any student encountered an item in each category. For example, in the constructed-response category, 37 students each saw at least one constructed-response item, and some saw more than one. This led to a total of 47 instances of students encountering constructed-response items.

Students mentioned that the answer options didn't reflect their understanding twice as frequently with a selected-response, single-prompt implicit multidimensional item, compared to a selected-response, multiple-prompt explicit multidimensional item. The issue would be impossible to occur with a constructed-response version of an item. Students mentioned feeling overwhelmed by the amount of information presented in the problem, including the number of answer options, more frequently when they used the selected-response multiple-prompt explicit multidimensional or selected-response single prompt implicit multidimensional item variations. Students were able to arrive at a "correct" answer by test-taking strategies, misunderstanding, or some other construct-irrelevant factor fairly equally across all item variations. Students reported rethinking an original answer upon seeing the answer options 13% and 15% of the time when a selected-response format was used. This would be impossible to observe with a constructed-response item. Finally, students infrequently misunderstood or misinterpreted the answer options, but this was not a problem with the constructed-response format.

The frequencies of the 3 sources of misunderstanding that are likely to be affected by scaffolding can be found in Table 4.3, for Round 2. Students misunderstood or misinterpreted the task more frequently on the selected-response versions of the task than on the constructed-response version. This may have been due to the selected-response version of the argument question, which allowed students to select more than one response. Many students asked for clarification about how to answer the argument question with a selected-response format. This was not the case for the constructed-response version of the argument. Students overlooked a critical piece of information

when all item formats were used, with no clear pattern of response format and/or scaffolding being related to this issue. As in Round 1, very few students expressed confusion about the item layout during this round of cognitive interviews.

Some students expressed a preference for one response format over another. Note that the sample of students who expressed a preference is nonrandom - most students were not asked to outright state a preference, so only students who spontaneously stated a preference or who displayed a behavior that prompted the researcher to ask them about

Table 4.3.

Frequency of Sources of Misunderstanding Affected by Scaffolding: Round 2

	Constructed response/Multiple prompt explicit multidimensional		Selected response/Multiple prompt explicit multidimensional		Selected response/Single prompt implicit multidimensional	
Issue	Count	Percentage	Count	Percentage	Count	Percentage
Misunderstanding or misinterpretation of critical piece of task	8	17.02%	21	30.00%	15	25.42%
Overlooks a critical piece of information	5	10.64%	3	4.29%	5	8.47%
Confused by item layout	2	4.26%	0	0.00%	0	0.00%
Total number of items viewed by students during Round 2*	47	100.00%	70	100.00%	59	100.00%

*Indicates the total number of times that any student encountered an item in each category. For example, in the constructed-response category, 37 students each saw at least one constructed-response item, and some saw more than one. This led to a total of 47 instances of students encountering constructed-response items.

their preference gave this information. Nevertheless, more students expressed a preference for the selected-response versions of the items (N=15) than for the constructed-response versions of the items (N=7).

Students who saw selected-response versions of the items were sometimes asked to complete an open-ended prompt before completing a selected-response version of the same prompt. For selected-response multiple-prompt explicit multidimensional versions of the items, students were asked to compose a written argument before they were asked to complete the selected-response version of the same prompt. For selected-response single-prompt implicit multidimensional versions of the items, students were asked to compose a written response before they completed the selected-response version of the same prompt.

The selected-response format seemed to influence student interactions with the item. As noted above, 13% and 15% of students were led to rethink their written answer after seeing the response options in the selected-response versions of the items. To further examine the role of response format in student response processes, their written and chosen responses were examined to establish the degree of concordance between responses that utilize different formats. Written responses and selected-responses were compared, and each pair was assigned 1 of 5 possible codes reflecting the degree to which the selected-responses matched the written response. These codes were: Match-plus, indicating that the selected-responses contained additional evidence and/or reasoning not present in the written response; Match, indicating that the selected-responses closely or exactly matched the written response; Match-minus, indicating that the written response contained additional evidence and/or reasoning not present in the

selected responses; No Match, indicating that the core reasoning in the written and selected-responses was different or contradictory; and No Written Response, indicating that no comparison was possible because the student chose selected-response options without providing a written response. The frequency of each match code can be found in Table 4.4.

Most commonly, students stuck with their original written argument even after they were presented with response options (77% and 82% of responses for selected-response multiple-prompt explicit multidimensional and selected-response single-prompt implicit multidimensionalitems, respectively). Many students provided additional evidence and/or reasoning on their selected-response arguments, when compared to their original written arguments (49% and 14% of responses). It was less common for students to change their argument upon seeing the response options (11% and 14% of responses).

Table 4.4.

Frequency of Matches Between Written and Selected-Response Arguments.

	Selected response/ Multiple prompt explicit multidimensional		Selected response/ Single prompt implicit multidimensional	
Match code	Frequency	Percentage	Frequency	Percentage
Match-plus	34	48.57%	8	13.56%
Match	17	24.29%	31	52.54%
Match-minus	3	4.29%	9	15.25%
No Match	8	11.43%	8	13.56%
No Written Response	8	11.43%	3	5.08%
Total Number of Responses*	70	100.00%	59	100.00%

*Indicates the total number of times that any student encountered an item in each category. For example, in the selected-response/multiple-prompt category, 37 students each saw at least one constructed-response item, and many saw more than one. This led to a total of 70 instances of students encountering selected-response/multiple-prompt items.

Some students declined to provide a written argument, but selected one or more response options on the selected-response version of the prompt.

In the selected-response version of a multiple-prompt multidimensional item, a written argument and selected-response argument were directly compared. For selected-response single-prompt implicit multidimensional versions of the items, if students' initial written answer did not include an argument, they were also asked to provide a written argument in support of their answer. Many students provided additional evidence or reasoning in their selected-response that was not present in their written arguments (49% of responses). Sometimes, additional selected-response evidence or reasoning did not support or was not relevant to their written argument. All written and selected-response arguments were examined, and the number of pieces of relevant-supporting evidence, irrelevant/unsupportive evidence, and reasoning were tabulated for each argument. The average frequency of reasoning and relevant-supporting evidence of written and selected-response arguments is found in Table 4.5. The cell values refer to the average amount of evidence and reasoning found in students' responses across the various item structures and response formats. For example, row 1 indicates that students provided 2.41 pieces of evidence, on average, to selected-response arguments, and 1.44-1.52 pieces of evidence, on average, in written arguments.

On average, students provided more evidence in selected-response arguments than written arguments, including more relevant-supporting evidence and more irrelevant/unsupportive evidence. Students also provided more reasoning in selected-response arguments. This is unsurprising, given that selected-response arguments allowed

Table 4.5.

Average Number of Pieces of Reasoning and Relevant-Supporting Evidence in Written and Selected-Response Arguments.[18]

	Multiple prompt explicit multidimensional		Single prompt implicit multidimensional
Argument Component	Selected-response Argument (SE)	Written Argument (SE)	Written Argument (SE)
Evidence	2.41 (0.28)	1.52 (0.15)	1.44 (0.16)
Relevant-supporting evidence	1.75 (0.20	1.24 (0.15)	1.31 (0.15)
Irrelevant-unsupporting evidence	0.65 (0.15)	0.27 (0.06)	0.16 (0.06)
Reasoning	1.62 (0.14)	0.84 (0.12)	0.97 (0.09)

students to choose among several potential pieces of evidence and reasoning, rather than evidence and reasoning from scratch as in the written arguments.

Number of dimensions addressed in student response. Student responses from the cognitive interviews and pilot tests were examined to compare the number of dimensions addressed in response to items with varying levels of multidimensional scaffolding and response formats.

Cognitive interviews. After the Round 1 cognitive interviews, student's written responses were examined to determine whether or not responses provided information about their understanding of the three assessment dimensions: Structure and Properties of Matter; Scale, Proportion, and Quantity, and Engaging in Argument from Evidence (NGSS Lead States, 2013). The written responses were classified as either providing or

[18] Note that the total number of observations comes from 10 item scenarios and 37 students. For each argument component, the average includes multiple observations from the same student on different items, and multiple observations from the same item with different students. This means that the observations are not independent. The standard errors have not been corrected to account for this dependence. The averages in this table should be interpreted cautiously.

not providing information relevant to assigning a score on these three dimensions. The number of dimensions addressed was counted for each response. If a student did not provide a written response, this was interpreted as addressing zero of the three assessment dimensions. (Students provided a written response in most cases – in 118 of 130 student responses. The 12 remaining item responses were distributed roughly equally across the three item variations.) The average number of dimensions addressed in responses across the three multidimensional scaffolding item variations is presented in Table 4.6.

Students addressed the highest number of dimensions, on average, when presented with multiple-prompt explicit multidimensional items and the fewest number of dimensions, on average, when presented with single-prompt implicit multidimensional items. To explore this pattern further, each dimension was examined individually. The

Table 4.6.

Average Number of Dimensions Addressed in Student Responses[19]

	Average number of dimensions addressed	Number of observations	Standard error
Multiple-prompt explicit multidimensional	2.65	43	0.12
Single-prompt explicit multidimensional	2.07	43	0.15
Single-prompt implicit multidimensional	1.57	44	0.12

Table 4.7.

Percentage of Written Responses Addressing Each Dimension

[19] Note that the total number of observations of each item variation comes from only 10 item scenarios and 26 students. For each of the item variations, the average number of dimensions addressed in a written response includes multiple observations from the same student on different items, and multiple observations from the same item with different students. This means that the observations are not independent. The standard errors have not been corrected to account for this dependence. The averages in this table should be interpreted cautiously.

	Structure and Properties of Matter	Scale, Proportion, and Quantity	Engaging in Argument from Evidence
Multiple-prompt explicit multidimensional	93.02%	81.40%	91.70%
Single-prompt explicit multidimensional	83.72%	48.84%	69.77%
Single-prompt implicit multidimensional	81.82%	31.82%	38.64%

percentage of responses within each item variation that addressed Structure and Properties of Matter; Scale, Proportion, and Quantity; and Engaging in Argument from Evidence can be found in Table 4.7.

Overall, students were more likely to address the Disciplinary Core Idea – Structure and Properties of Matter – in their responses than any other dimension. This was consistent across the three item variations. Students were most likely to address each of the three dimensions with the multiple-prompt explicit multidimensional item, followed by the single-prompt explicit multidimensional item, and least likely to address each dimension when presented with single-prompt implicit multidimensional items.

In particular, the Scale, Proportion, and Quantity dimension was most likely to be neglected when the single-prompt formats were utilized. This pattern might be manifest due to the nature of these assessment tasks; the Structure and Properties of Matter subtask tended to be the most salient focus of the item scenarios, whereas the Scale, Proportion, and Quantity subtask tended to be an intermediate step towards completing the Structure and Properties of Matter subtask. The Engaging in Argument from Evidence subtask was the written justification for the student's response to the other two subtasks. When a single-prompt format was used, it was difficult to write assessment items in such a way

that avoided highlighting some of the dimensions and sidelining others, which affects the observed relationship between the dimensions. Even though the three dimensions are treated as conceptually independent entities, they are dependent in the single-prompt assessment context. This may account for the observed pattern, in which the Scale, Proportion, and Quantity dimension is addressed only 32% and 49% of the time in the single-prompt implicit and explicit multidimensional cases, respectively. The multiple-prompt format seems to somewhat mitigate student inattention to the Scale, Proportion, and Quantity dimension, and in fact all three dimensions see increases from the single-prompt explicit multidimensional format to the multiple-prompt format.

Also note that the Engaging in Argument from Evidence dimension was explicitly cued in the single-prompt explicit multidimensional versions of the items. In these versions of the items, the second part of the response stem asks students to provide an argument including evidence and reasoning. This statement may be considered a second prompt within the single-prompt format, although students were only provided with one response space (whereas the multiple-prompt format provided them with a separate response space for each dimension). This may explain why the Engaging in Argument Dimension tends to have a higher response rate when the single-prompt explicit structure is used, compared to the single-prompt implicit structure. The Scale, Proportion, and Quantity dimension is not explicitly cued in either single-prompt structure, which may explain why it is the most neglected dimension among both single-prompt item variations.

Assessment pilot. The number of dimensions addressed by students was examined again with the pilot data, this time by looking at the frequency of missing data among

students' scored responses. Tables 4.8, 4.9, and 4.10 display the amount of missing data from the assessment pilot, broken apart by item type (amount of multidimensional scaffolding and response format) and dimension (Scale, Proportion and Quantity, Structure and Properties of Matter, and Engaging in Argument from Evidence). Responses were scored as "Blank on All Dimensions" if the student did not provide any response to the entire item context. For a single-prompt item, this occurred when the response area was completely blank. For a multiple-prompt item, this occurred when the response areas for each of the individual sub-prompts were *all* completely blank. Responses were scored as "Missing on [Dimension]" if the student provided a response that addressed part, but not all of the item context. For a single-prompt item, this occurred when a student provided a response, but did not address all of the assessment dimensions. For example, a student could provide an argument defending their understanding of the matter concept from the item, neglecting the Scale, Proportion, and Quantity aspect of the item context. In this case, the student would receive a score of "Missing on Scale, Proportion, and Quantity Dimension" and would receive the appropriate score for the remaining two dimensions according to the rubric. For a multiple-prompt item, this occurred when a student responded to at least one, but not all of the item's dimensional sub-prompts. The blank sub-prompts received a score of "Missing on [Dimension]" and sub-prompts with responses were rated according to the scoring rubric.

Table 4.8.

Frequency of Missing and Blank Data on the Scale, Proportion, and Quantity Dimension

	Multiple-prompt		Single-prompt		Constructed-response		Selected-response	
	N	Percentage	N	Percentage	N	Percentage	N	Percentage
Blank on All Dimensions	72	9.90%	129	17.11%	9	6.08%	10	6.49%
Partially Blank/Missing on SPQ Dimension	54	7.43%	134	17.77%	1	0.68%	1	0.65%
Scored	601	82.67%	498	65.12%	138	93.24%	143	92.86%
Total	727	100.00%	754	100.00%	148	100.00%	154	100.00%

Table 4.9.

Frequency of Missing and Blank Data on the Structure and Properties of Matter Dimension

	Multiple-prompt		Single-prompt		Constructed-response		Selected-response	
	N	Percentage	N	Percentage	N	Percentage	N	Percentage
Blank on All Dimensions	72	9.90%	129	17.11%	9	6.08%	10	6.49%
Partially Blank/Missing on Matter Dimension	34	4.68%	93	12.33%	12	8.11%	4	2.60%
Scored	621	85.42%	539	70.56%	127	85.81%	140	90.91%
Total	727	100.00%	753	100.00%	148	100.00%	154	100.00%

Table 4.10.

Frequency of Missing and Blank Data on the Engaging in Argument from Evidence Dimension

	Multiple-prompt		Single-prompt		Constructed-response		Selected-response	
	N	Percentage	N	Percentage	N	Percentage	N	Percentage
Blank on All Dimensions	72	9.90%	129	17.11%	61	10.25%	62	10.65%
Partially Blank/Missing on Argument Dimension	65	8.94%	51	6.76%	56	9.41%	16	2.75%
Scored	590	81.16%	574	76.13%	478	80.34%	504	86.60%
Total	727	100.00%	753	100.00%	595	100.00%	582	100.00%

On the Scale, Proportion, and Quantity dimension, more missing responses were observed on single-prompt versions of the items than on multiple-prompt versions of the items. Cumulatively, the gain of information on multiple-prompt versions of the items was substantial; about 20% more student responses addressed the Scale, Proportion, and Quantity dimension when a multiple-prompt format was used. Many more responses were classified as missing when a single-prompt version of the item was used, compared to the multiple-prompt version.

A chi-squared test of independence was conducted to examine the association between the number of prompts (single or multiple) and a student's response status (Blank on All Dimensions, Missing on Scale, Proportion, and Quantity Dimension, or scoreable). Results of the chi-squared test are not interpreted in terms of significance, as the data come from a non-random sample of items and students, and there is no larger population of inference; however, the results are presented as guidelines to indicate the presence of potentially interesting patterns. There was an association between the number of prompts and a student's response status ($\chi^2 = 50.46$, df = 2)[20]. Standardized residuals reveal that there are more missing responses (of both types) than expected when the single-prompt format is used, and fewer missing responses than expected when the multiple-prompt format is used.

There was no observed association between response format (constructed- or selected-response) and student response status ($\chi^2 = 0.022$, df = 2).

[20] Note that each cell in the table contains aggregated data from various items and students, and data from the same student may be present in more than one cell, thus violating the assumption that observations are independent. Results should be interpreted cautiously.

On the Structure and Properties of Matter dimension, more missing responses (of both types) were observed on single-prompt versions of the items than on multiple-prompt versions of the items. Although the loss of information was not as large as on the Scale, Proportion, and Quantity dimension, the differences are still striking. About 15% more students provided scoreable information about the Structure and Properties of Matter dimension when a multiple-prompt version of the item was used, compared to the single-prompt version. Many more responses were classified as missing when a single-prompt version of the item was used, compared to the multiple-prompt version.

A chi-squared test of independence revealed an association between the number of prompts (single or multiple) and a student's response status (Blank on All Dimensions, Missing on Structure and Properties of Matter Dimension, or scoreable) ($\chi^2 = 24.32$, df = 2)[21]. Again, these results are presented as guidelines and not as indicators of significance. Standardized residuals reveal that there are more responses that were classified as Blank on All Dimensions than expected when the single-prompt format is used, and fewer responses that were classified as Blank on All Dimensions than expected when the multiple-prompt format is used.

Based on the percentages in Table 4.9, students were slightly more likely to provide scoreable information on selected-response version of the items than constructed-response versions of the items, with a difference of about 5% between the two response

[21] Note that each cell in the table contains aggregated data from various items and students, and data from the same student may be present in more than one cell, thus violating the assumption that observations are independent. Results should be interpreted cautiously.

formats.[22] A chi-square test of independence reveals no association between response status and response format ($\chi^2 = 4.57$, df = 2).

On the Engaging in Argument from Evidence dimension, Table 4.10 shows that differences between multiple-prompt and single-prompt versions of the items were small. On multiple-prompt versions of the items, slightly more responses were classified as Missing on the Engaging in Argument from Evidence Dimension than on single-prompt versions of the items. The direction of the difference is notable, because it is opposite of the difference observed on the other two dimensions.

A chi-squared test of independence revealed an association between the number of prompts (single or multiple) and a student's response status (Blank on All Dimensions, Missing on the Engaging in Argument from Evidence Dimension, or scoreable) ($\chi^2 = 17.587$, df = 2) [23]. Again, these results are presented as guidelines and not as indicators of significance. Standardized residuals reveal that there are more responses classified as Blank on All Dimensions than expected when the single-prompt format is used, and fewer responses were classified as Blank on All Dimensions than expected when the multiple-prompt format is used. Note that the difference in the expected and observed frequencies of responses scored as Missing on the Engaging in Argument from Evidence Dimension is small; in this case, the number of prompts does not seem to affect the likelihood that students will address the argument dimension, given that they provide any response at all. This may be due to the fact that single-prompt items tended to emphasize

[22] Note that the total sample size for this comparison is small, given that only one item directly compared response formats on the Structure and Properties of Matter dimension.
[23] Note that each cell in the table contains aggregated data from various items and students, and data from the same student may be present in more than one cell, thus violating the assumption that observations are independent. Results should be interpreted cautiously.

the Engaging in Argument from Evidence dimension by explicitly cuing students to provide their evidence and reasoning.

The difference between constructed-response and selected-response items was a bit larger – from the table, there is a roughly 7% increase in responses classified as Missing on the Engaging in Argument from Evidence Dimension on constructed-response items compared to selected-response items. An exploratory Chi-squared test revealed an association ($\chi^2 = 22.703$, df = 2) [24] between response format and student response status, although this result is intended for guidance rather than for interpretation of statistical significance.

Overall, it appears that the number of response prompts has a relationship with the frequency of responses classified as Missing on [Dimension] and Blank on All Dimensions, such that there are more of both types of missing responses on single-prompt versions of the items. The constructed-response format also seems to be related to the frequency of both types of missing responses on the Engaging in Argument from Evidence dimension.

Impact of Multidimensional Scaffolding on Students of Different Abilities. Student response data for the assessment pilot included 64 items from three dimensions (22 Scale, Proportion, and Quantity items; 20 Structure and Properties of Matter items; and 22 Engaging in Argument from Evidence items). Student response data was scaled using ConQuest 4 (Adams, Wu, & Wilson, 2015). Marginal maximum likelihood estimation was used to generate item difficulty estimates and fit statistics for each dimension, and

[24] Note that each cell in the table contains aggregated data from various items and students, and data from the same student may be present in more than one cell, thus violating the assumption that observations are independent. Results should be interpreted cautiously.

WLE person ability estimates for all three dimensions. Based on WLE person ability estimates, the sample was divided into three approximately equal subgroups of high, medium, and low ability students on each dimension. Within each subgroup, the extent of missing data was examined again. The percentage of missing data was calculated by counting the number of responses scored as Missing on [Dimension] across all items of a particular type (i.e., all single-prompt items, or all multiple-prompt items) and dividing by the total number of items of that type that were administered across all students.

As previously noted, responses scored as "Missing on [Dimension]" were more frequent when the single-prompt item format was used, with the exception of the Engaging in Argument from Evidence dimension, where the reverse pattern is observed. This pattern is observed across students of all ability levels. Furthermore, lower ability students tended to have more responses scored as "Missing on [Dimension]" than high ability students, especially when a single-prompt format was used. The gap between the multiple-prompt and single-prompt formats is largest among low ability students. This gap is still present among medium and high ability students, but it is smaller in size (Figure 4.4). On the Engaging in Argument from Evidence dimension, the direction of the difference was reversed, and the size of the gap was smaller among all ability groups.

Table 4.11.

Percentage of Responses Scored as Missing on [Dimension]

	Scale, Proportion, and Quantity		Structure and Properties of Matter		Engaging in Argument from Evidence	
Subgroup	Multiple-prompt	Single-prompt	Multiple-prompt	Single-prompt	Multiple-prompt	Single-prompt
Low ability	8.70%	23.02%	3.63%	16.00%	10.89%	6.15%
Medium ability	9.47%	15.00%	4.15%	5.98%	7.85%	2.06%
High ability	7.38%	12.24%	2.63%	5.26%	4.41%	2.66%

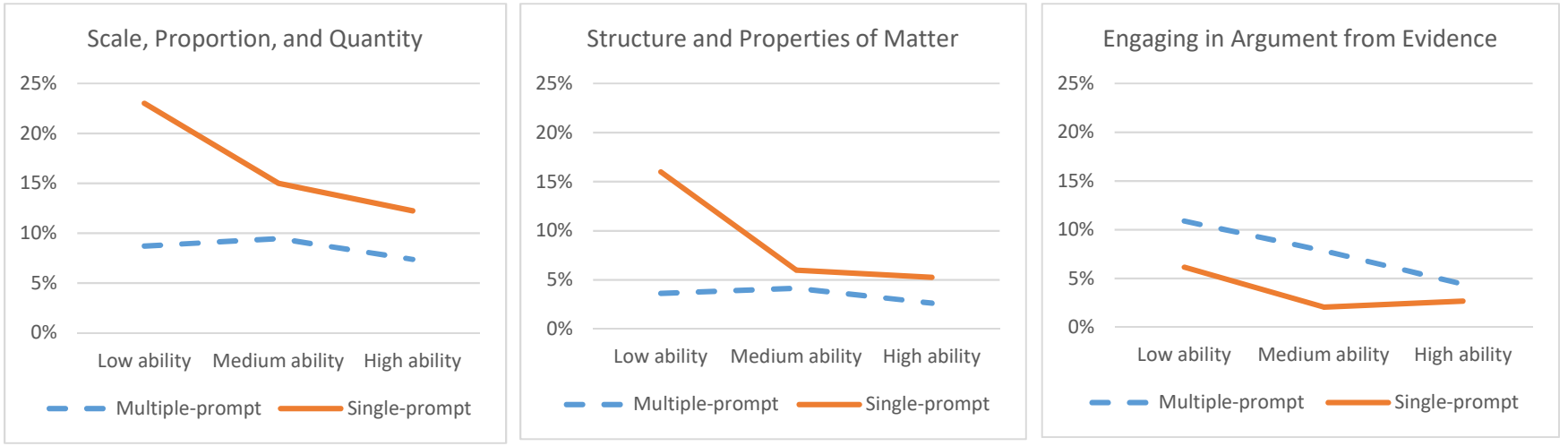


Figure 4.4. Percentage of responses scored as "Missing on [Dimension]."

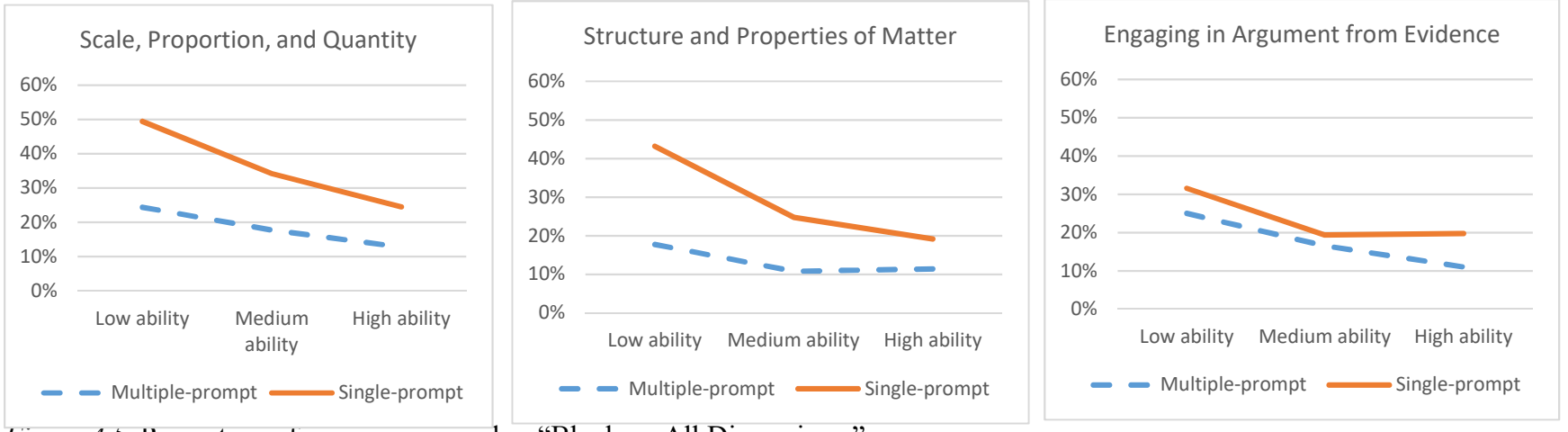


Figure 4.5. Percentage of responses scored as "Blank on All Dimensions."

These findings suggest that the effect of scaffolding largely depends on student ability. Low-ability students are particularly impacted by the amount of scaffolding in an assessment item, and are more likely to provide an answer that includes evidence of their ability on all dimensions when scaffolding is present. On the Engaging in Argument from Evidence dimension, however, the pattern is obscured. The nature of a single-prompt item encourages an extended response by explicitly emphasizing that students should provide extended reasoning about their thinking. In this case, the multiple-prompt format does not provide extra emphasis and structure for the Engaging in Argument from Evidence dimension in the same way that it does for the other two dimensions, which may explain why the difference between the multidimensional scaffolding variations is smaller on this dimension.

The differences between single and multiple prompt items were even more extensive among those responses scored as Blank on All Dimensions (Table 4.12). On the Scale, Proportion, and Quantity and Structure and Properties of Matter dimensions, low ability students are the most likely to provide a response that is Blank on All Dimensions (e.g., forego a response altogether), regardless of item format. Yet across all ability levels, students tend to provide more responses that are classified as Blank on All Dimensions to single-prompt items, and the gap between single-prompt and multiple-prompt items decreases as student ability increases, on both dimensions – a striking finding that is discussed further in Chapter 5. However, the gap is still present even among high ability students.

Table 4.12.

Percentage of Responses Scored as Blank on All Dimensions

	Scale, Proportion, and Quantity		Structure and Properties of Matter		Engaging in Argument from Evidence	
Subgroup	Multiple-prompt	Single-prompt	Multiple-prompt	Single-prompt	Multiple-prompt	Single-prompt
Low ability	24.35%	49.43%	17.74%	43.20%	25.00%	31.56%
Medium ability	17.70%	34.17%	10.79%	24.79%	16.53%	19.34%
High ability	12.70%	24.49%	11.40%	19.17%	11.01%	19.77%

On the Engaging in Argument from Evidence dimension, the overall gap between single and multiple prompt items is much smaller (Figure 4.5), potentially because of the nature of single-prompt items described in the previous pages.

Response time. During both rounds of the cognitive interviews, response time was recorded and aggregated for each multidimensional scaffolding and response format variation. Although all students were asked to read and complete each item in full before beginning their interview, some students ignored this request and asked the researcher for clarification about the item as they were still responding. Response times for these students are skewed, as they include time spent in conversation with the researcher in addition to time spent answering the questions. These response times were removed from the dataset. In addition, some students were unable to produce a written response at all, and no measure of response time is available for these cases.

Multidimensional scaffolding. The average response time was calculated with the remaining responses (80% of all possible responses, total N = 130), and a summary can be found in Table 4.13.[25]

[25] Note that the total number of observations comes from 10 item scenarios and 26 students. For each scaffolding variation, the average includes multiple observations from the same student on different items, and multiple observations from the same item with different students. This means that the observations are

Table 4.13.

Average Response Time for Multidimensional Scaffolding Item Variations

Multidimensional scaffolding variation	Average response time	Standard error	N
Multiple-prompt explicit multidimensional	3:19	0:17	36
Single-prompt explicit multidimensional	2:22	0:17	32
Single-prompt implicit multidimensional	1:32	0:10	36

Response time increased monotonically with the amount of item scaffolding. Single-prompt implicit multidimensional items tended to be completed more quickly, averaging just a minute and a half from start to finish. Multiple-prompt explicit multidimensional items took the longest time to complete – more than twice the amount of time as a single-prompt implicit multidimensional item, on average. Single-prompt explicit multidimensional items, which required only a single response but cued students to attend to all 3 dimensions of the assessment, split the difference between the other scaffolding variations.

It is possible that these averages underestimate the true amount of time that it would take for students to answer the items, because missing data mainly occurred when students were confused about some aspect of the task, and elected to ask the researcher instead of continuing. These students would likely have taken more time to wrestle with their confusion in a real testing environment. Their absence from the dataset may bias the computed averages.

Response format. During Round 2 of the cognitive interviews, response time was recorded and aggregated for each item variation. In the case of selected-response

not independent. The standard errors have not been corrected to account for this dependence. The averages in this table should be interpreted cautiously.

variations, students were given both selected-response and constructed-response versions of some prompts. This was done to enable direct comparisons between student responses under each format. For these items, the various item components were given to students separately and sequentially, such that students were always asked to complete a constructed-response version of an item prompt before answering the selected-response version. This ensured that their exposure to distractors did not affect comparisons between responses to selected- and constructed-response prompts. On the selected-response versions of multiple-prompt explicit multidimensional items, students were first given the selected-response Structure and Properties of Matter and Scale, Proportion, and Quantity prompts, and a constructed-response Engaging in Argument from Evidence prompt. Upon completion of these prompts, students were given the selected-response version of the Engaging in Argument from Evidence prompt. On the selected-response versions of the single-prompt implicit multidimensional items, students were first given a single-prompt constructed response prompt, followed by a selected-response prompt, and finally a constructed-response argument prompt (only administered if their initial response did not address the Engaging in Argument from Evidence dimension). Response time was recorded for the written and selected-response components of the item separately. Again, cases were removed from the dataset when the student interrupted or gave no written response. The average response time was calculated with the remaining responses, and a summary can be found in Table 4.14.

Table 4.14.

Average Response Time for Response Format Variations.[26,27]

Response format variation	Item component	Average response time	Standard error	N
Constructed-response/ Multiple-prompt explicit multidimensional	Whole item	3:05	0:12	47
Selected-response/ Multiple-prompt explicit multidimensional	Cumulative SPQ, matter, and constructed-response argument prompts	3:00	0:11	68
	Constructed-response argument only	1:17	0:04	100
	Selected-response argument only	0:54	0:05	59
Selected-response/ Single-prompt implicit multidimensional	Constructed-response/Single-prompt implicit multidimensional	1:33	0:10	79
	Selected-response/Single-prompt implicit multidimensional	0:27	0:03	49
	Constructed-response argument only	1:16	0:07	56

Replacing the Structure and Properties of Matter and Scale, Proportion, and Quantity prompts with selected-response versions of the same prompt did not substantially affect response time; it took students only 5 seconds less, on average, to respond to a multiple-prompt explicit multidimensional item where only the argument required a constructed-response. The response time for selected-response and constructed-response arguments differed by about 23 seconds, on average, with the

[26] Sample sizes are larger than the total number of responses for each variation because four items required multiple explanations. Response time for each explanation was used as a separate observation in the dataset.

[27] Note that the total number of observations comes from 10 item scenarios and 37 students. For each response format variation, the average includes multiple observations from the same student on different items, and multiple observations from the same item with different students. This means that the observations are not independent. The standard errors have not been corrected to account for this dependence. The averages in this table should be interpreted cautiously.

constructed-response version of that prompt requiring more response time. It is difficult to compare the response times of single-prompt implicit multidimensional items with different response formats, because reading time is included in the average time for the constructed-response version, but not the selected-response version. Therefore, the best information about response time comes from comparing the constructed-response and selected-response versions of the multiple-prompt multidimensional items, and suggests that students take 28 seconds less per item, on average, to respond to selected-response prompts about all three dimensions.

Interrater reliability. After the pilot test, interrater reliability was examined by calculating an intraclass correlation coefficient (ICC) from a two-way random-effects analysis of variance. Both "consistency" and "absolute" measures were calculated for each score. Consistency measures take into account the covariance in a group of ratings, even when the raters do not reach complete agreement, whereas absolute measures account for only absolute agreement between raters.

ICC's were calculated for each item based on the scores from both the holistic and multidimensional rubric. See Appendix C for ICC's of all items for both rubrics. Table 4.19 contains information about the average ICC's for each item variation, computed from raters' scores using the multidimensional rubric. Table 4.20 contains information about the average ICC's for each item variation, computed from raters' scores using the holistic rubric. Consistency measures are used in both tables. ICC's were averaged by first using Fisher's Z transformation, averaging the resulting z-scores, and then reverse-transforming back to ICC's.

Table 4.15.

Average ICC's for Ratings for Each Item Variation Using a Multidimensional Rubric

Item variation	SPQ	Matter	Argument
Constructed-response multiple-prompt explicit multidimensional	0.99	0.99	0.61
Constructed-response single-prompt explicit multidimensional	0.90	0.99	0.52
Selected-response multiple-prompt explicit multidimensional	0.99	0.99	0.73

Table 4.16.

Average ICC's for Ratings for Each Item Variation Using a Holistic Rubric

Item variation	ICC
Constructed-response multiple-prompt explicit multidimensional	0.75
Constructed-response single-prompt explicit multidimensional	0.83
Selected-response multiple-prompt explicit multidimensional	0.72

Multidimensional scaffolding. Note from Table 4.15 that interrater reliability tended to be very high for multiple-prompt items on the Scale, Proportion, and Quantity and Structure and Properties of Matter dimensions. When this structure was used, student responses to the Scale, Proportion, and Quantity and Structure and Properties of Matter dimensional sub-prompts tended to be brief and straightforward; consequently, the scoring rubrics were also straightforward, and the raters seemed to distinguish between the scoring categories with ease. Interrater reliability was high on the Structure and Properties of Matter dimension, regardless of item structure or response format. Raters seemed to make distinctions about students' understanding of the matter concepts with ease.

When using the multidimensional rubric, interrater reliability tends to be higher for multiple-prompt items than for single-prompt items, when response format is held

constant. This difference in reliability is observed on two of the three dimensions: Scale, Proportion, and Quantity and Engaging in Argument from Evidence. There is no observable difference in interrater reliability among the item variations on the Structure and Properties of Matter dimension; all variations have very high interrater reliability on this dimension. When using the holistic rubric (Table 4.16), interrater reliability is higher for single-prompt items than for multiple-prompt items. This is the reverse of the observed pattern under the analytic/multidimensional rubric, and suggests that interrater reliability varies depending on an interaction between scaffolding and rubric structure.

Response format. For the multidimensional rubric, there was no difference in the average interrater reliability of constructed-response and selected-response items from the Scale, Proportion, and Quantity and Matter dimensions. In the argument scoring category, selected-response items demonstrated higher rates of interrater reliability, on average, when the amount of multidimensional scaffolding is held constant. For the holistic rubric, the average interrater reliability of selected-response items was slightly lower than that of constructed-response items, when the amount of multidimensional scaffolding is held constant.

Variation in interrater reliability for students of different abilities. As in the missing data analysis, the multidimensional dataset was split into three approximately equal groups based on WLE person ability estimates. Due to the split, the sample size of double-scored responses decreased substantially, ranging from 10 to 40 depending on the item. Thus, the resulting reliability estimates may not be representative of their true values in the population. Intraclass correlation coefficients (ICC's) were used as the primary indicator of interrater reliability for each item. ICC's were transformed via

Fisher's z-transformation, and then combined to indicate average reliability for 4 item variations (single- and multiple-prompt; constructed- and selected-response) among low, medium, and high ability students. Due to the small sample sizes and few items in each variation, statistical significance tests were not performed on ICC's. Furthermore, the Scale, Proportion, and Quantity, and Structure and Properties of Matter dimensions were not included in this analysis, due to having demonstrated robust interrater reliability estimates in the overall analysis.

Looking at patterns in reliability among students of different ability, reliability tended to be highest among medium ability students (Table 4.17), and lower among the low and high ability groups. When data is limited to the subset of 5 item scenarios with single- and multiple-prompt item variations, this pattern holds across both scaffolding variations (Figure 4.6). The middle scoring categories may be easier for raters to apply consistently, or the middle scoring categories may become "default" when student responses are difficult to categorize according to the rubric. When data is limited to the subset of 4 item scenarios with constructed- and selected-response format variations

Table 4.17.

Average ICC's (Consistency Measure) for Ratings Among Student Ability Groups for All Items, Argument Dimension.

	ICC
Low	0.48
Medium	0.55
High	0.38

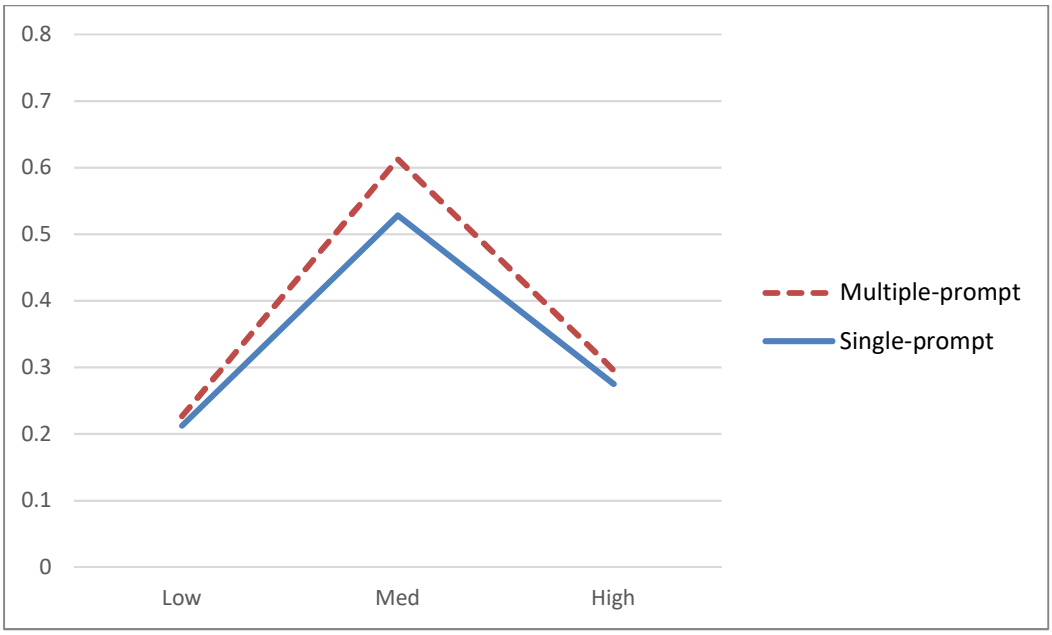


Figure 4.6. Average ICC's for Multidimensional Scaffolding Item Variants for Low, Medium, and High Ability Students, Argument Dimension.

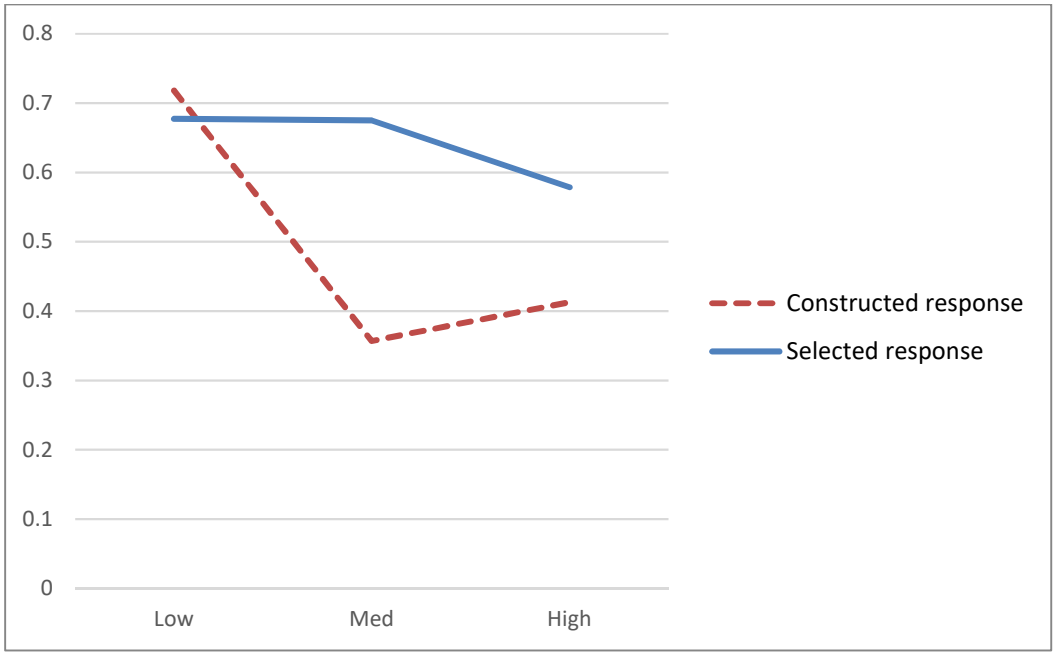


Figure 4.7. Average ICC's for Ratings for Response Format Item Variants for Low, Medium, and High Ability Students, Argument Dimension.

(Figure 4.7), there was some fluctuation in interrater reliability depending on response format. Interrater reliability tended to be more stable across ability groups when a selected-response format was used, whereas the ICC fluctuated in size across ability groups when a constructed-response format was used. The items in Figures 4.6 and 4.7 are mutually exclusive subsets of items, however, the constructed-response arguments in Figure 4.7 are very similar in structure and format to the multiple-prompt items in Figure 4.6 (albeit with different content). Yet the pattern of ICC's across ability groups is considerably different for the multiple-prompt items in Figure 4.6 compared to the constructed-response items in Figure 4.7. This may be due to chance variation in reliability estimates due to small sample size, or item-specific variations. Regardless, there seems to be substantial variation among ability groups in the single- prompt, multiple-prompt, and constructed-response items, but less variation in the ICC among selected-response items. This suggests that the selected-response rubrics may have been easier to consistently apply at all scoring categories. The constructed-response rubrics (including single-and multiple-prompt items) may have been harder to apply, and differences in applying particular scoring categories may have been affected by item or rater specific factors.

Missing data and interrater reliability. When the multidimensional scoring rubric was used, raters were given the option to assign a score of Missing on [Dimension] to a particular item and dimension. For multiple-prompt items, this occurred when a student left one or more dimensional sub-prompts blank, but provided an answer to at least one of the sub-prompts within an item context. For single-prompt items, this occurred when a student provided a response that did not provide any information about their ability on

one or more of the assessment dimensions. Raters were instructed to mark a response as Blank on All Dimensions if the student did not respond to any part of the prompt/sub-prompts. Thus, raters' scores involved a series of judgments: first whether the student response contained information relevant to the dimension being scored, and second, assigning a score (if enough evidence existed). Discrepancies between raters may arise from either of these two judgments.

The intraclass correlations reported previously do not take into account any discrepancies between raters on whether or not a student's response is Missing on [Dimension]. This is because this classification is a categorical value, and therefore cannot be summarized with the numerical codes used for scoring. However, raters' judgments about whether or not a student's response was present or missing was an additional source of discrepancies among raters. Across all items and all students in the final dataset, about 5% of student responses were coded as "Missing on [Dimension]." Among the subset of responses that were scored by 2 raters[28], 64% of those responses were not scored as "Missing on [Dimension]" by both raters. This indicates that judgment of whether or not enough evidence exists to assign a score is a difficult one. Since single-prompt items displayed a larger frequency of "Missing on [Dimension]" responses (see Tables 4.8 through 4.10 starting on page 151), discrepancies in this judgment are especially likely to further weaken the interrater reliability of these items. Multiple-prompt and selected-response items exhibited less missing data and are therefore less likely to be affected.

[28] A subset of responses were double-scored to check interrater reliability, and student responses without missing information were prioritized for inclusion in the interrater sample. Therefore, the number of discrepancies observed in the interrater sample due to missing data may not be representative of the extent of discrepancies among all responses.

Reconciling rater unreliability. Of the 64 items (22 Engaging in Argument from Evidence items, 20 Structure and Properties of Matter items, and 22 Scale, Proportion, and Quantity items) that were administered during the pilot test, several items demonstrated low interrater reliability[29]. Based on the dataset from the multidimensional rubric, 28 of the 64 items had ICC's lower than 0.85, including all 22 Engaging in Argument from Evidence items, 2 Structure and Properties of Matter items, and 4 Scale, Proportion, and Quantity items. Low interrater reliability was a concern because it indicated that raters may have faced difficulty in interpreting and applying the scoring rubric. This was especially apparent for the Engaging in Argument from Evidence dimension, as all 22 items were characterized by low interrater reliability regardless of response format. The two affected Structure and Properties of Matter items and four Scale, Proportion, and Quantity items were some of the most difficult on the assessment, and assessed complex, hard-to measure skills. It seems likely that the task complexity may have played a role in raters' difficulty with the scoring process, in addition to items' scoring rubrics.

There are several approaches for reconciling rater discrepancies, including taking the mean of rater scores, soliciting a third rater or expert rater to determine the appropriate score, or asking the original raters to discuss and reach a consensus (Penny & Johnson, 2011). Furthermore, there are psychometric models that allow for separation and examination of the different factors (facets) that influence item performance, including rater effects and item difficulty; these are called multifaceted models (Myford & Wolfe, 2003).

[29] ICC's for each item using both the holistic and multidimensional rubrics are found in Appendix C.

Rater discrepancies on the 22 Engaging in Argument from Evidence items were frequent and widespread across all items. Due to limited resources, it was not feasible to employ reconciliation methods that involved additional raters or discussion between the original raters. Therefore, a multifaceted Rasch model was utilized to account for differences between raters. The multifaceted Rasch model extends the partial credit model (Chapter 3, page 81) by adding a new term C_r which is added to the item and step difficulty term δ_{ik}. Thus, the probability of a student's response becomes a function of ability θ_n, rater severity, and item/step difficulty (Linacre, 1989).

$$P(X_{ni} = x_i|\theta_n, \delta_{ik}) = \frac{\exp \sum_{k=0}^{x_i}(\theta_n - \delta_{ik} - C_r)}{\sum_{j=0}^{m_i} \exp[\sum_{k=0}^{j}(\theta_n - \delta_{ik} - C_r)]}$$

Using a multifaceted model, rater effects are estimated independently from the item and step difficulty estimates. All estimates are measured on the same scale, enabling comparisons between the various facets (Myford & Wolfe, 2003). Most importantly, the confounding effect of rater discrepancies is separated from the item estimates, meaning that the item difficulty estimates are more validly attributable to the items alone.

On the Structure and Properties of Matter and Scale, Proportion, and Quantity dimensions, only a few items demonstrated low interrater reliability. Items with lower interrater reliability tended to measure the same difficult skill: integration of concepts and computations related to proportions and material properties. Because of the inherent difficulty of this concept for upper elementary grades, some students demonstrated nuanced partial understandings that were difficult to score. Therefore, an expert rater was selected to provide the final judgment for all of these item responses[30].

[30] The expert rater in question was the author of this dissertation.

Before examining the dimensionality of students' scored responses, a baseline model was established for each dimension. On the Structure and Properties of Matter and Scale, Proportion, and Quantity dimensions, a partial credit model was fitted to each set of items. On the Argumentation dimension, the baseline model was established after comparing the fit of several multifaceted rating scale and partial credit models, including many variations with different rater effects and interactions. These models included 1) a rater model, which included parameters indicating the relative harshness/leniency of the 7 raters; 2) a rater by item interaction model, which included extra parameters to account for variation in the raters' behavior specific to a particular item; and 3) a model including rater by item and rater by item by step interactions, in which extra parameters were included to account for differences in how the raters assigned polytomous categories on particular items. Model fit was evaluated by comparing the difference in the deviance statistic (G^2) of hierarchical models to a Chi-squared distribution (Adams, Wilson, & Wang, 1997) (Table 4.18). All comparisons between hierarchical models were significant, indicating that a) in every case, the partial credit model fit the data significantly better than the rating scale model, and b) the greater the number of rater terms and interactions in the model, the better the fit. The best fitting model was the most complex model: a partial credit model including rater effects, rater by item interactions, and rater by item by step interactions. This model, which was used as the baseline model for the Engaging in Argument from Evidence items in the subsequent dimensionality analysis, predicts the log odds of a student's response to an item as a linear combination of the overall item difficulty, the difficulty associated with moving from response category k to response category k + 1 for a particular item, the harshness or leniency of

Table 4.18.

Comparison of Multifaceted Argument Models with Different Rater Effects and Interactions

Rating Scale Models:		
	Deviance (G^2)	Number of parameters
item + step (simple rating scale model)	9180.94	24
item + step + rater	9065.49	30
item + step + rater + item x rater	8954.27	46
item + step + rater + rater x step	8964.20	42
item + step + rater + item x rater + rater x step	8848.59	58

Partial Credit Models:		
	Deviance (G^2)	Number of parameters
item + item x step (simple partial credit model)	8926.48	66
item + rater + item x step	8796.69	72
item + rater + item x step + item x rater	8694.98	88
item + rater + item x step + item x rater + item x step x rater	8526.96	132

the particular rater who scored the response, the interaction between a particular item and rater, and the interaction between a particular rater, item, and the difficulty of moving from response category k to response category k + 1 on that item.

Item difficulty. Student performance data for all 64 items was scaled using ConQuest 4 (Adams, Wu, & Wilson, 2015). Marginal maximum likelihood estimation was used to generate item difficulty estimates and fit statistics for each dimension, and to estimate dimensional variances and covariances based on a multidimensional model. Item difficulty estimates and fit statistics for all items may be found in Appendix D, and the items themselves can be found in Appendix F.

Multidimensional scaffolding. Figures 4.8, 4.9, and 4.10 contain Wright maps for each assessment dimension, based on estimates from the multidimensional model. Wright

maps are a visual representation of estimated item difficulty and person ability. Easier items and lower achieving persons are located at the lower end of the figure, and harder items and higher achieving persons are located at the top. The Wright maps in Figures 4.8, 4.9, and 4.10 contains 5 pairs of items. Each pair shares a common task context, but one item (ITEM A) utilized a multiple-prompt format and the other (ITEM B) utilized a single-prompt format.

On the Scale, Proportion, and Quantity dimension, the direction of the differences in difficulty between multiple- and single-prompt items was not consistent for all item pairs. The multiple-prompt version of the item was easier for 2 of the 4 pairs of items (CAROL, ANA), and the single-prompt version of the item was easier for 2 of the 4 pairs (ROMITA, SUGAR). There doesn't appear to be any systematic advantage to using a multiple-prompt format, compared to a single-prompt.

On the Structure and Properties of Matter dimension, a pattern emerges. On average, single-prompt item thresholds are easier than their corresponding multiple-prompt item thresholds. This trend holds across all item pairs - for 4 of the 5 item pairs (ANA, SUGAR, CAROL, and BOX), the single-prompt version of the item is easier than the multiple-prompt version. For the remaining item (ROMITA), the items have roughly equivalent difficulty estimates. Thus, the single-prompt item format tends to be easier than the multiple-prompt format on this dimension. The single-prompt format may allow students to provide a response that emphasizes the strengths of their understanding and deemphasizes their weaknesses. Splitting the item into multiple responses may make it more difficult by forcing students to directly address misconceptions they may have masked in a single response. In addition, the single prompt format may encourage a halo

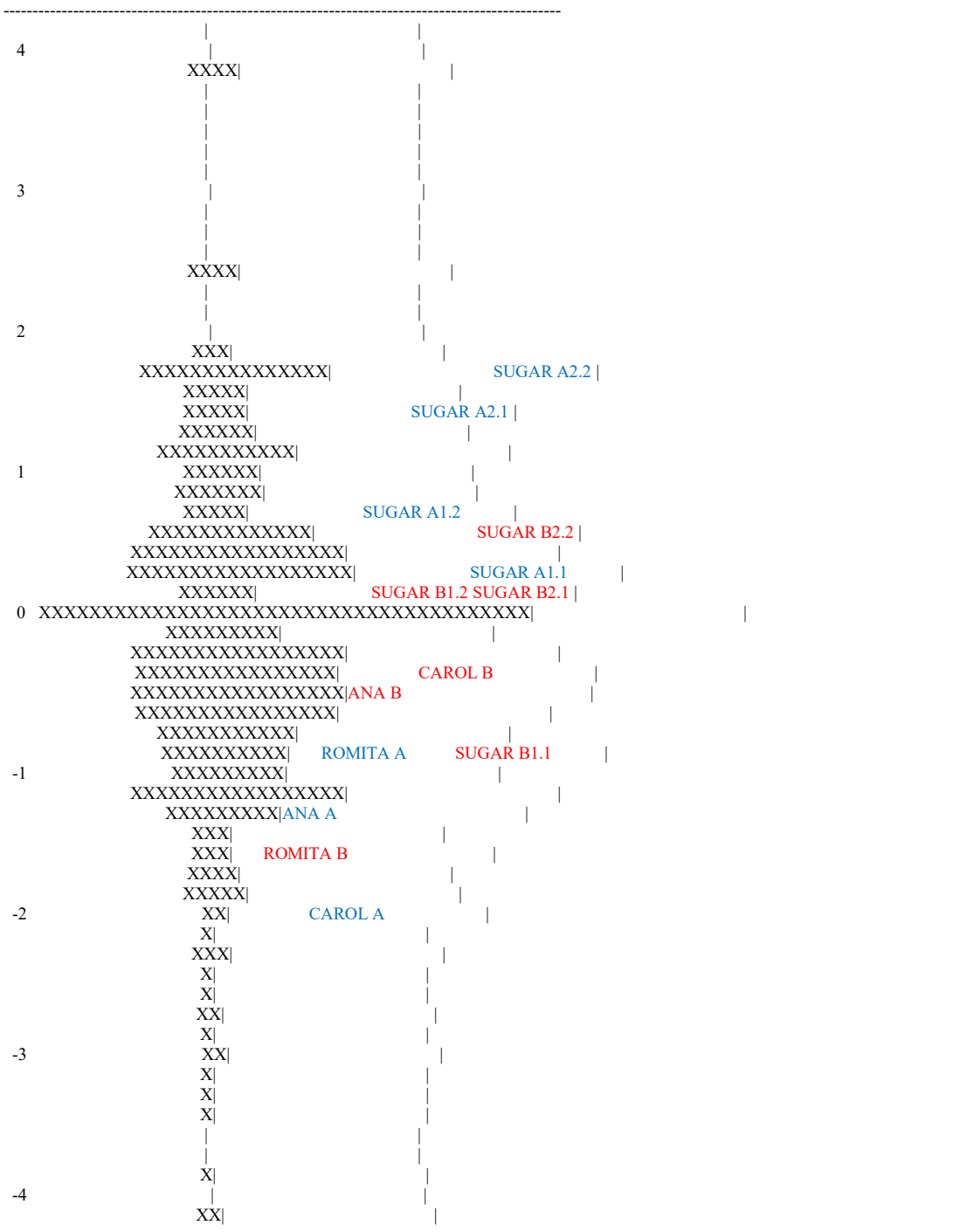


Figure 4.8. Item difficulty comparison of different multidimensional scaffolding variations on the Scale, Proportion, and Quantity dimension. Version A items have a multiple-prompt format, while Version B items have a single-prompt format. [31]

[31] Only 4 items are included in the Scale, Proportion, and Quantity Wright map for multidimensional scaffolding. One of the items, labeled BOX in the Matter and Argument Wright maps, was scored differently on the Scale, Proportion, and Quantity dimension, depending on whether it had multiple-prompts or a single-prompt. Therefore, the item difficulty estimates for the two variants are not directly comparable on this dimension.

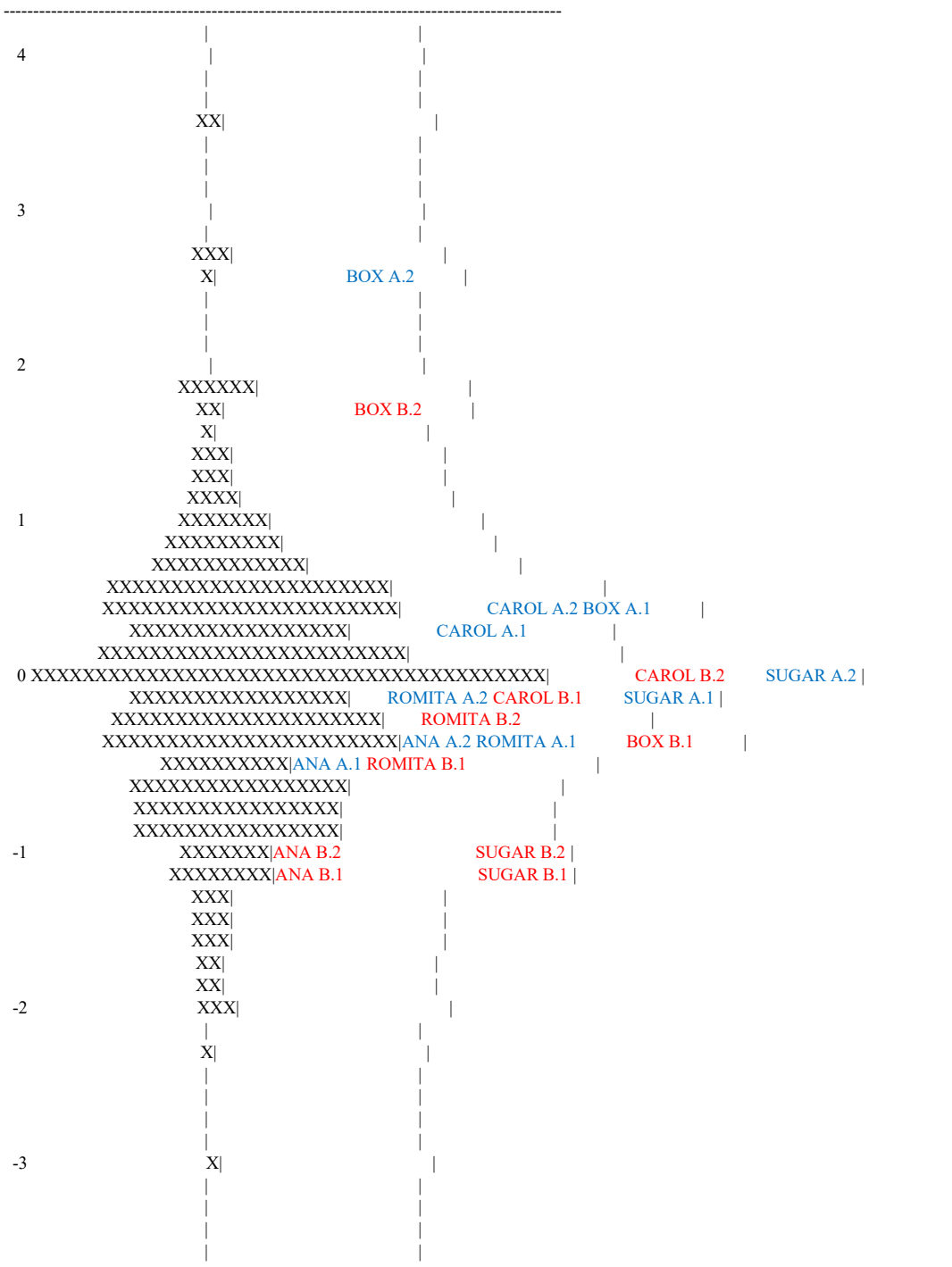


Figure 4.9. Item difficulty comparison of different multidimensional scaffolding variations on the Structure and Properties of Matter dimension. Version A items have a multiple-prompt format, while Version B items have a single-prompt format.

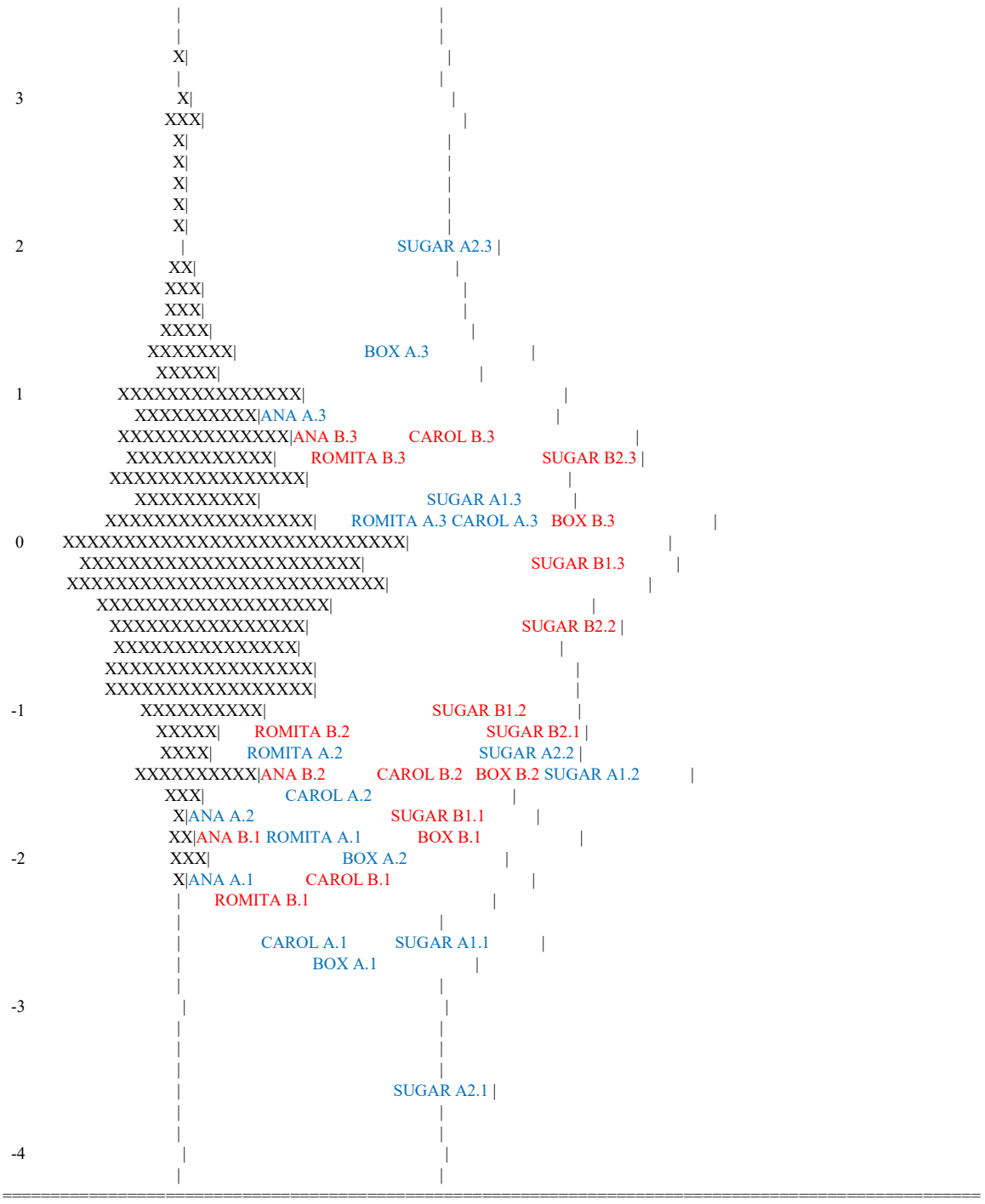


Figure 4.10. Item difficulty comparison of different multidimensional scaffolding variations on the Engaging in Argument from Evidence dimension. Version A items have a multiple-prompt format, while Version B items have a single-prompt format.

effect among raters. A longer discussion of these possible reasons for the observed difference in item difficulty can be found in Chapter 5.

On the Engaging in Argument from Evidence dimension, there is a slight tendency for single-prompt thresholds to cluster near the center of the scale distribution while the multiple-prompt thresholds have a larger range. This results in a pattern unique to the Engaging in Argument from Evidence dimension where multiple-prompt items tend to be easier than single-prompt items at the lower scoring categories and harder than single-prompt items at the higher scoring categories. There are many factors which may contribute to this pattern, including rater conflation of the dimensions (a halo effect), differences in student behavior to single prompt and multiple prompt items, or chance. These possibilities are discussed in more depth in Chapter 5.

Response format. There were 4 pairs of items that utilized both response format variations. All 8 items used a multiple-prompt format, but one item from each pair utilized a selected-response format and the other utilized a constructed-response format. Of these 4 pairs, 3 of them varied the response format for the Argument prompt only – an identical response format was used for the Scale, Proportion, and Quantity, and Matter prompts (usually selected-response, but sometimes a mixed response format). Only one item (KEVIN) varied the response format on all three prompts. Therefore, Wright maps were only examined for the Engaging in Argument from Evidence dimension (Figure 4.11).

At the lower thresholds, there is no clear pattern in item difficulty among selected-response and constructed-response argument items. In some cases (i.e., KEVIN, NATE), the selected-response threshold is easier than the corresponding constructed-

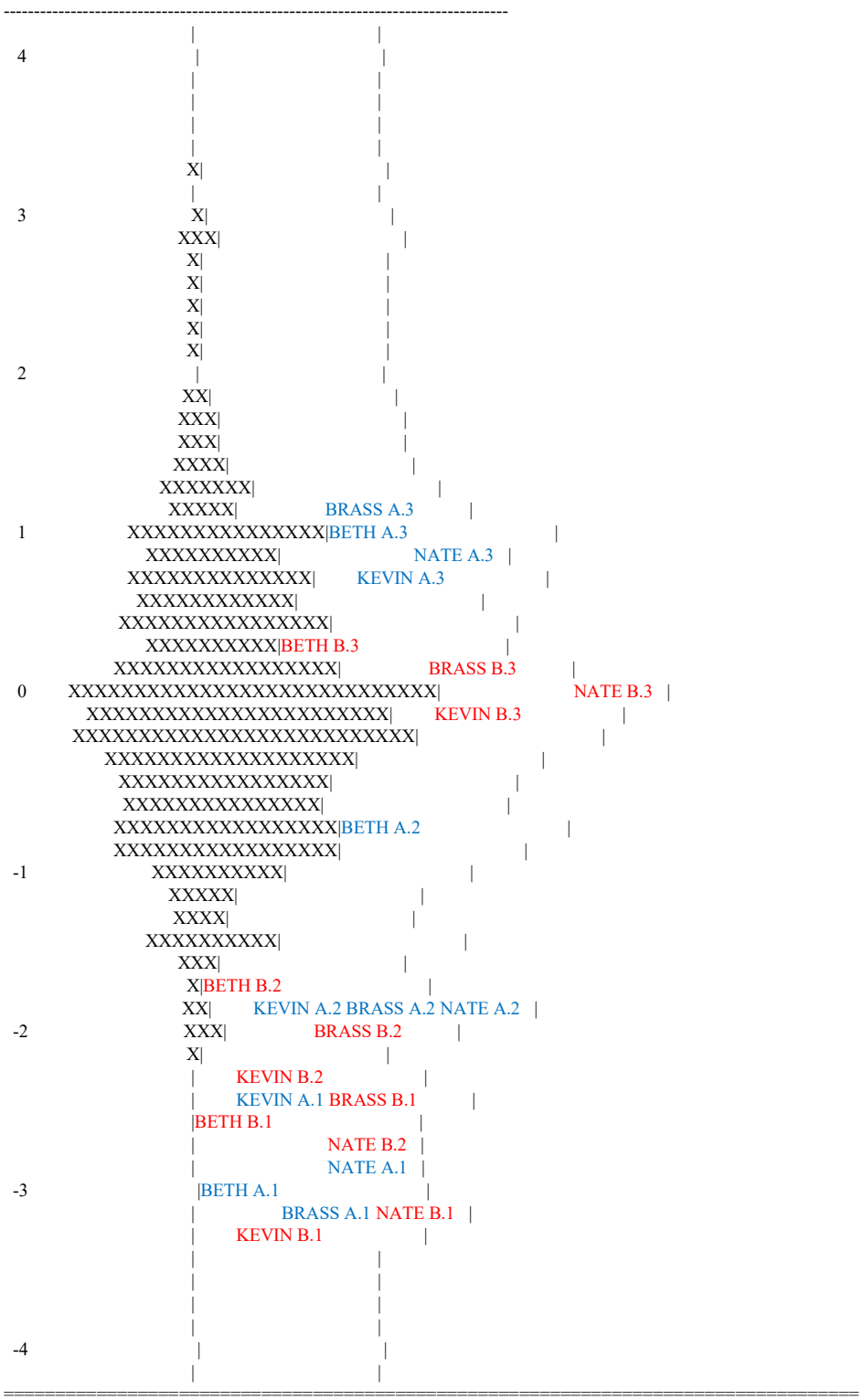


Figure 4.11. Item difficulty comparison of different response format variations on the Engaging in Argument from Evidence dimension. Version A items have a constructed-response format, while Version B items have a selected-response format.

response threshold. In other cases (i.e., BETH, BRASS), the constructed-response threshold is easier. However, at the third threshold, the multiple-choice threshold is systematically easier than the constructed- response threshold, across all items. When it comes to providing the most sophisticated arguments, which include both evidence and reasoning, it appears that the multiple-choice format offers students an advantage. At the lower thresholds, no advantage is apparent based on item format. Furthermore, there is no disadvantage associated with the multiple-choice format at the lower levels, suggesting that the increased reading load required by multiple-choice items does not seem to pose a barrier to entry for students. This finding contradicts conventional wisdom about the selected-response format, where distracters are generally seen as a potential source of construct-irrelevant variance due to additional reading comprehension demands when compared to constructed-response items. The observed pattern casts doubt on these assumptions, although the role of sample characteristics (e.g., reading proficiency, English language proficiency) cannot be ruled out as a potential explanation.

Differences in item difficulty associated with multidimensional scaffolding among students of different abilities. When the analysis was repeated with three subgroups defined by student ability on each of the assessment dimensions, some results held while others became muddled. Analysis focused on the differences between item variations, and whether the size and direction of these differences vary with students of different abilities. The item difficulty estimates here were calculated based on small sample sizes (often smaller than N=30), and are not suitable for precise comparisons; however, they are presented here for descriptive purposes. As a rough guideline, a

difference larger than 0.5 logits is often considered substantial enough to merit examination in subgroup analyses (Agustin, 2006).

On the Scale, Proportion, and Quantity dimension, the size and direction of differences between item variations is largely consistent across all ability groups. As in the overall analysis, there is no clear pattern in the differences between single-prompt and multiple-prompt items.

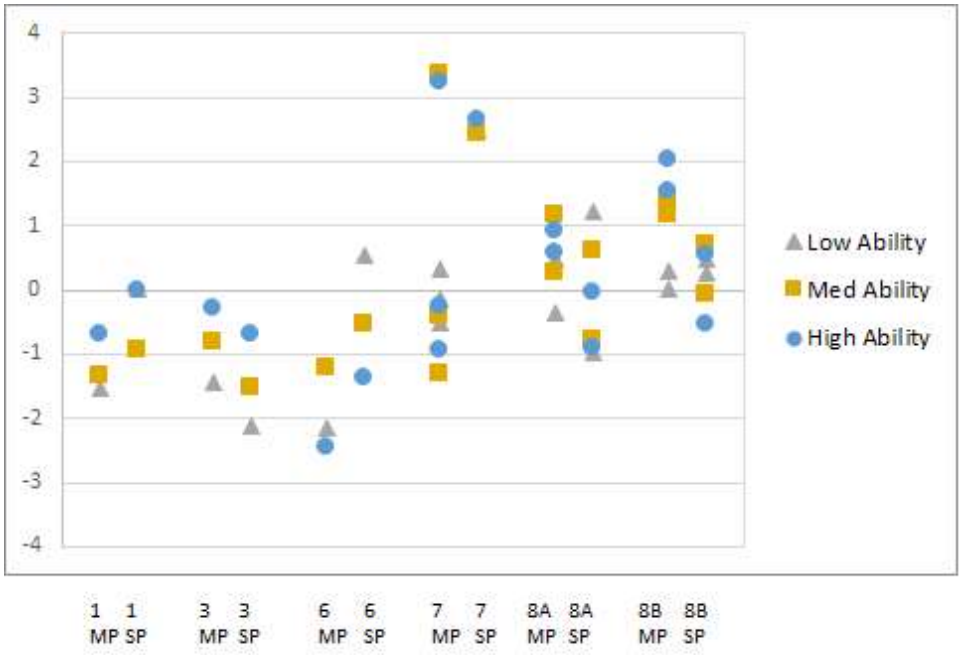


Figure 4.12. Item difficulty maps for multiple-prompt and single-prompt items on the Scale, Proportion, and Quantity dimension when students are split into subgroups based on ability. Item variants with identical context but different types of scaffolding (i.e., single-prompt, multiple-prompt) are presented in pairs along the X-axis. MP indicates the multiple-prompt version, and SP indicates the single-prompt version. The Y-axis indicates item difficulty, in logits. Points on the lower end of the scale indicate easier items/thresholds, and points near the top indicate more difficult items/thresholds.

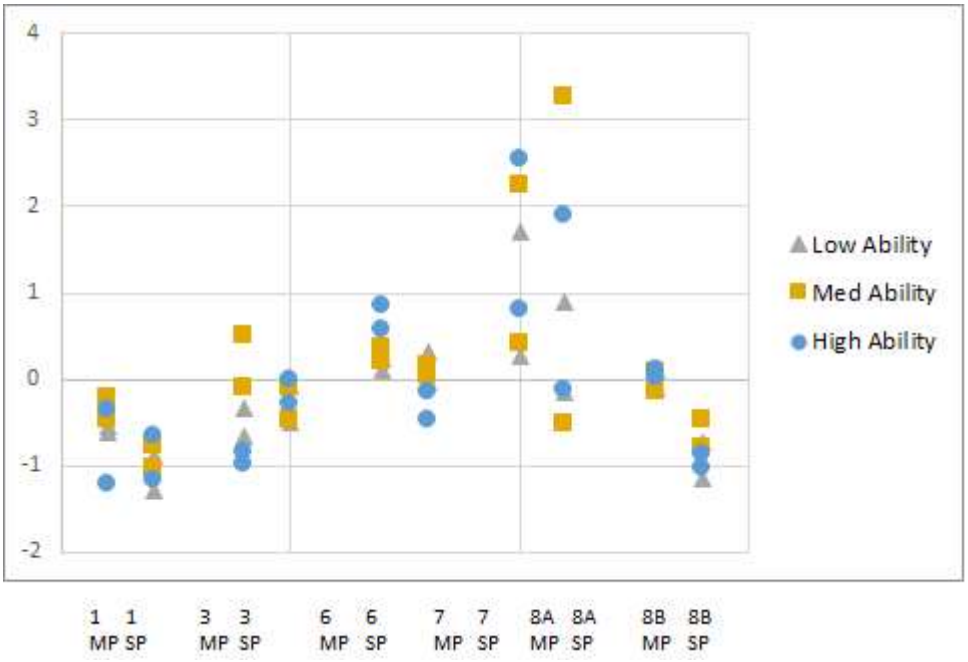


Figure 4.13. Item difficulty maps for multiple-prompt and single-prompt items on the Structure and Properties of Matter dimension when students are split into subgroups based on ability. Item variants with identical context but different types of scaffolding (i.e., single-prompt, multiple-prompt) are presented in pairs along the X-axis. MP indicates the multiple-prompt version, and SP indicates the single-prompt version. The Y-axis indicates item difficulty, in logits. Points on the lower end of the scale indicate easier items/thresholds, and points near the top indicate more difficult items/thresholds.

On the Structure and Properties of Matter dimension, the overall pattern that single-prompt items tend to be easier than multiple-prompt items holds across all ability groups. The one exception to this pattern is item 3, which also deviated in the overall analysis.

On the Engaging in Argument from Evidence dimension, patterns in item difficulty appear to be fairly constant across all 3 ability groups. The lower two thresholds are sometimes easier when the single-prompt format is used, and often the difficulty is very similar across the two item variations. But the top threshold is

systematically easier when the single-prompt format is used, and this holds across all ability groups.

The previously observed overall pattern that single-prompt thresholds tend to be clustered is inconsistent, and in some cases reverses once the dataset is split into the three ability groups. Without a clear pattern, it's difficult to speculate about why differences in difficulty between multiple-prompt and single-prompt items may occur.

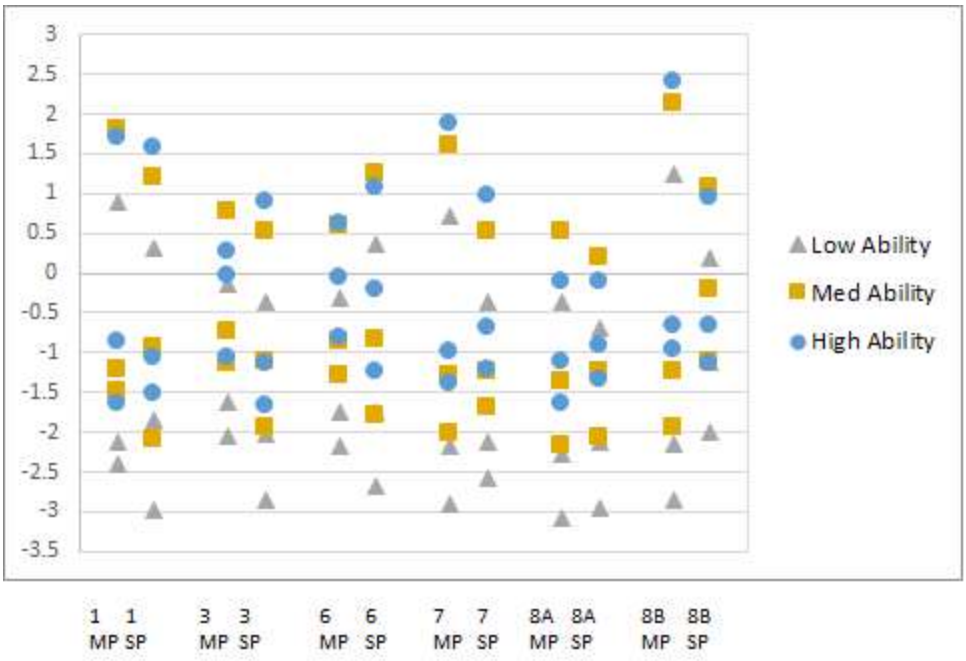


Figure 4.14. Item difficulty maps for multiple-prompt and single-prompt items on the Engaging in Argument from Evidence dimension when students are split into subgroups based on ability. Item variants with identical context but different types of scaffolding (i.e., single-prompt, multiple-prompt) are presented in pairs along the X-axis. MP indicates the multiple-prompt version, and SP indicates the single-prompt version. The Y-axis indicates item difficulty, in logits. Points on the lower end of the scale indicate easier items/thresholds, and points near the top indicate more difficult items/thresholds.

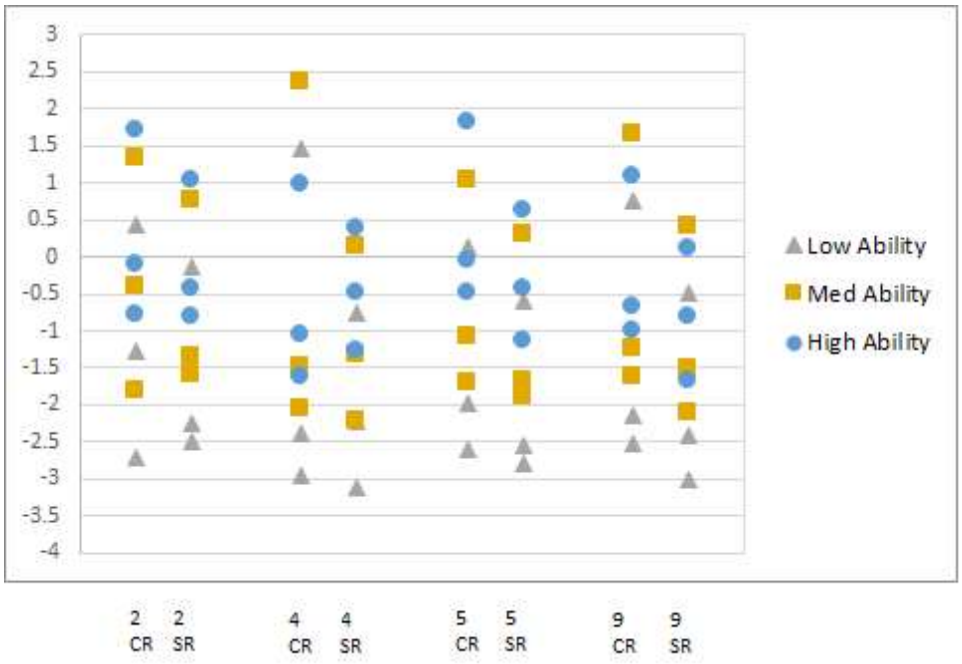


Figure 4.15. Item difficulty maps for constructed-response and selected-response items on the Engaging in Argument from Evidence dimension when students are split into subgroups based on ability. Item variants with identical context but different types of scaffolding (i.e., single-prompt, multiple-prompt) are presented in pairs along the X-axis. CR indicates the constructed-response version, and SR indicates the selected-response version. The Y-axis indicates item difficulty, in logits. Points on the lower end of the scale indicate easier items/thresholds, and points near the top indicate more difficult items/thresholds.

Differences in item difficulty associated with response format among students of different abilities. On the Engaging in Argument from Evidence dimension, the selected-response top thresholds are easier than constructed-response across all ability groups. This mirrors the pattern observed with the overall group, and reinforces the hypothesis that the selected-response format changes the nature of the task.

There is still no discernible pattern of difficulty in the bottom thresholds for selected-response or constructed-response items, again suggesting that reading load does not pose as big of a barrier as commonly supposed.

Item misfit. All 64 items (22 Scale, Proportion, and Quantity items, 20 Structure and Properties of Matter items, and 22 Engaging in Argument from Evidence items) were examined for misfit using a weighted mean square statistic. Marginal maximum likelihood estimation was used to generate item difficulty estimates and fit statistics for each dimension. The weighted mean square is an indicator of the degree to which student responses to an item conform to the expected responses. The weighted mean square statistic comes with an associated t-statistic. Large values of the weighted mean square (greater than 1.33), or t-values greater than 1.96 were flagged as indicators of misfit, indicating that students' responses did not conform to the model. Small values of the weighted mean square (less than 0.75), or t-values less than -1.96 may indicate that students' responses conform to the model to a higher degree than anticipated. Item difficulty estimates and fit statistics for all items may be found in Appendix D.

On the Scale, Proportion, and Quantity dimension, no items were flagged as misfitting based on the listed criteria. There is no indication that response format or multidimensional scaffolding either improve or detract from an item's fit to the model on this dimension.

On the Structure and Properties of Matter dimension, two items were flagged as having high positive misfit. For one of these (BRASS A), the source of item misfit is unclear. This item utilized both a multiple-prompt structure and a selected-response format, but its comparison item (BRASS B) utilized the same structure and response format for the particular sub-prompt that was flagged. Both items had similar fit statistics (BRASS A weighted MNSQ = 1.15, t=1.6; BRASS B weighted MNSQ = 1.21, t=2.2), although only BRASS B was greater than the stated threshold. The cause of any potential

misfit is likely attributable to a common element of the two items, and may or may not involve the response format.

The second item flagged on the Structure and Properties of Matter dimension was KEVIN A, an open-ended multiple prompt item, which was flagged as having significant positive misfit (weighted MNSQ = 1.25, t = 2.30). Its selected-response counterpart (KEVIN B) fit the model well (MNSQ = 1.09, t = 1). This suggests that the difference in response format is a potential source for misfit. The item asked students to calculate the weight of several coffee beans by subtracting the weight of the container holding the beans. Then, students were asked to predict how much the container of coffee would weigh after the beans were ground into powder. According to rater reports, many students' responses indicated that they were unsure of whether to account for the weight of the container when making their prediction, and many provided responses that describe the weight of the coffee without the container. In the selected-response version of the item, the response options mitigated this dilemma, since all options were relative to the initial weight of the container and beans (e.g., "A little bit more than X grams", "A little bit less than X grams", "X grams", etc.). In this case, it appears that the selected-response options made the task clearer for students and likely improved the item's fit.

On the Engaging in Argument from Evidence dimension, 12 of the 22 items were flagged as misfitting based on t-statistics higher than 1.96. Focusing on only the subset of items that have multidimensional scaffolding variations, there is no clear relationship between multidimensional scaffolding and misfit. Of four pairs of item variations, two of the multiple-prompt variations were flagged for misfit, while the corresponding single-

prompt variations fit well, and two of the remaining single-prompt variations were flagged for misfit, while their corresponding multiple-prompt variations fit well.

Looking at the selected-response variations on the Engaging in Argument from Evidence dimension, a pattern emerges. In two cases, the selected-response version of the item was flagged for misfit, while the constructed-response version of the same item was not. For the remaining two item pairs, both selected-response and constructed-response variations were flagged for misfit. In all four pairs, the value of the selected-response mean square was higher than the corresponding value of the constructed-response mean square. It appears that the selected-response format may be related to higher incidence of misfit when used to assess student proficiency with argumentation. This is likely because the selected-response format allows some students to engage in different response processes; i.e., some students may be forming an argument and then selecting the response(s) that best capture their original argument, while others may evaluate and select response options without first forming an argument of their own. The latter response process is unique to the selected-response variation, as students must form their own arguments without response options on the constructed-response version. This additional potential response process may also be responsible for an observed difference in difficulty between the selected- and constructed-response variations: at the highest thresholds, providing a complete scientific argument based on both evidence and reasoning was more difficult when a constructed-response format was used. Other factors that may influence the observed misfit include the high reading load in selected-response arguments, variable student interpretation of the response options, raters' differential application of the Engaging in Argument from Evidence scoring rubric with selected-

response items, or item-specific factors. Further discussion of these factors may be found in Chapter 5.

Variation in the extent of item misfit for students of different abilities. The dataset was again split into 3 ability groups, and the three subgroups were examined for patterns in misfitting items. The Scale, Proportion, and Quantity and Structure and Properties of Matter dimensions did not demonstrate any misfitting items at any ability level. As before, the extent of misfit is substantial on the Engaging in Argument from Evidence dimension for both overall item parameter estimates and step parameter estimates. Group means for low, medium, and high ability students were compared across item variations (Table 4.19). Differences between group means were not evaluated for statistical significance, as the data come from a non-random sample of items and students, and there is no larger population of inference; however, patterns in average item misfit are interpreted descriptively. Selected-response items tended to have more extreme fit statistics than constructed-response items, across all ability groups. This indicates more unpredictable responses when a selected-response format was used, and supports the pattern observed in the overall dataset. There are no clear patterns in average misfit

Table 4.19.

Average Misfit T-statistics for Multiple- and Single-Prompt, Selected- and Constructed-Response Argument Items Across Low, Medium, and High Ability Groups

	Low Ability	Medium Ability	High Ability
Constructed-response	1.05	0.72	1.2
Selected-response	2.01*	2.33*	1.66
Multiple-prompt	0.59	1.40	0.84
Single-prompt	0.91	1.49	1.03

*Indicates an average misfit T-statistic outside the acceptable range from -1.96 to 1.96.

among the ability groups, when selected-response and constructed-response items are compared.

When multiple-prompt and single-prompt items are compared, medium ability students tended to have more extreme misfit statistics than low ability students. This indicated that students whose abilities were close to the mean tended to have more unpredictable responses on the group of single- and multiple-prompt items included in this analysis, while high and low ability students were more likely to conform to expectation. This suggests that very weak and very strong students tend to interact with the items more predictably, perhaps as a result of their extreme ability. Multiple-prompt items tended to have slightly better fit than single-prompt items across all ability groups; however, the size of the difference is small and fluctuates across ability groups.

Research question 3: To what extent do unidimensional and multidimensional scoring and modelling approaches affect the relationships among the 3 dimensions of science learning (assuming that such relationships exist)?

The NGSS proposes a three-dimensional model of science learning. To examine whether the assessment aligned with the NGSS's theoretical structure, and to examine the impact of the scoring rubric and psychometric model on the assessment's dimensionality, three models were compared:

1. Model 1: A unidimensional Rasch model, combined with a multidimensional (a.k.a. analytic) scoring rubric,
2. Model 2: A three dimensional Rasch model, where the dimensions are defined in accordance with the NGSS, combined with a multidimensional (a.k.a. analytic) scoring rubric, and
3. Model 3: A unidimensional Rasch model, combined with a holistic scoring rubric.

Several psychometric criteria are examined for each of the three models:

Model deviance. Deviance statistics for the unidimensional and multidimensional models from the analytic dataset can be found in Table 4.20. These models are hierarchical models, and therefore the difference in model deviance can be directly compared to a Chi-squared distribution (Adams, Wilson, & Wang, 1997). Note that the holistic model is also included in this table, although it cannot be directly compared to the other two models since it comes from a different set of scores and many fewer items. The difference between the unidimensional and multidimensional model was statistically significant (χ^2 difference = 109.76, df = 5, $p < 0.01$). The multidimensional model demonstrates significantly better fit to the data than the unidimensional model.

Table 4.20.

Model Deviance from the Unidimensional and Multidimensional Models

	Deviance (G^2)	Number of Parameters
Analytic Unidimensional	16460.50	204
Analytic Multidimensional	16370.48	209
Holistic Unidimensional	9516.79	121

Item fit. Item fit statistics illustrate the differences in overall model deviance in finer detail. All item estimates and fit statistics can be found in Appendix D. Using the multidimensional dataset, there were a total of 64 items. Of the 22 Engaging in Argument from Evidence items, 14 were flagged as misfitting when a unidimensional model was employed and the multidimensional dataset was used, compared to 11 items flagged when a multidimensional model was employed. All of the Scale, Proportion, and Quantity items fit the scale well under both models. Two Structure and Properties of Matter items demonstrated significant positive misfit, and this occurred with both the unidimensional and the multidimensional model. The Argument items fit the scale slightly better under the multidimensional model; this explains why the multidimensional model demonstrated significantly better overall fit compared to the unidimensional model.

When a holistic rubric was used (in conjunction with a unidimensional model), 15 of the 20 items were flagged as misfitting. Overall, it appears that a multidimensional model paired with a multidimensional rubric provides the best item fit, and a unidimensional model paired with a holistic rubric provides the worst item fit.

Correlation between dimensions. Correlations between the three dimensions, estimated from the multidimensional model, are found in Table 4.21. Correlations are fairly high, ranging from about 0.7 to 0.9. The strongest correlation is between the Scale,

Proportion, and Quantity and Structure and Properties of Matter dimensions at 0.89. These correlations are quite high, indicating a strong degree of concordance between the dimensional person estimates. This somewhat contrasts the model deviance results from Table 4.20; although the difference between models is statistically significant, the correlations suggest that the distinction between dimensions is not significant in a practical sense. In general, an ability estimate's relative ranking on one dimension implies a similar relative ranking on another dimension. Keep in mind, however, that all items shared a narrow content area, which may inflate the interdimensional correlations. Depending on the quality of other psychometric indicators (such as reliability), it may be justifiable to sacrifice some nuance in interpretation in exchange for more robust statistical characteristics of the model.

The Engaging in Argument from Evidence dimension demonstrated the lowest correlation with the other two dimensions, and it shows the largest improvement in item fit with the multidimensional structure (above). This dimension seems to be the most different from the others, such that multidimensional structure will provide the greatest amount of additional nuance in the interpretation of student estimates.

To examine variation in student performance across dimensions, multidimensional WLE estimates were converted to z-scores and compared across dimensions. Based on these z-scores, about 50% of respondents have ability estimates on one of the dimensions that differ from estimates on at least one other dimension by a standard deviation or more. The largest number of discrepancies occur between the Scale, Proportion, and Quantity and Engaging in Argument from Evidence dimensions (31% of students). The remaining dimensional pairs each had about 25% of students with

Table 4.21.

Correlations Between the Three Dimensions: Scale, Proportion, and Quantity; Structure and Properties of Matter; and Engaging in Argument from Evidence

	Structure and Properties of Matter	Engaging in Argument from Evidence
Scale, Proportion, and Quantity	0.89	0.73
Structure and Properties of Matter	--	0.80

discrepant estimates. The scatterplots in Figure 4.16 further illustrate these discrepancies. The Scale, Proportion, and Quantity vs. Engaging in Argument from Evidence scatterplot has the largest amount of variation around the trend line, with 20% of variation in students' argumentation estimate predicted by their Scale, Proportion, and Quantity Estimate. The remaining dimensional pairs also demonstrate some variation around the trend line, to a lesser degree. Although the relationships between subscale ability estimates are strong, inconsistencies in student performance across dimensions are widespread. Multidimensional ability estimates would provide some useful information to tease apart student performance in these cases.

The relationship between ability estimates on the Scale, Proportion, and Quantity and Structure and Properties of Matter dimensions, in particular, is very strong. To test whether these dimensions might be better represented by a collapsed scale, a two-dimensional model was fitted. Engaging in Argument from Evidence was left as a separate dimension, but all remaining items were scaled together. Overall model deviance was 16388.21, with 206 estimated parameters. A chi-squared test revealed that the original three-dimensional model fit the data significantly better than the collapsed two-dimensional model (χ^2 difference = 17.73, df = 3, $p < 0.01$). However, the two-dimensional model had significantly better fit than the unidimensional model (χ^2

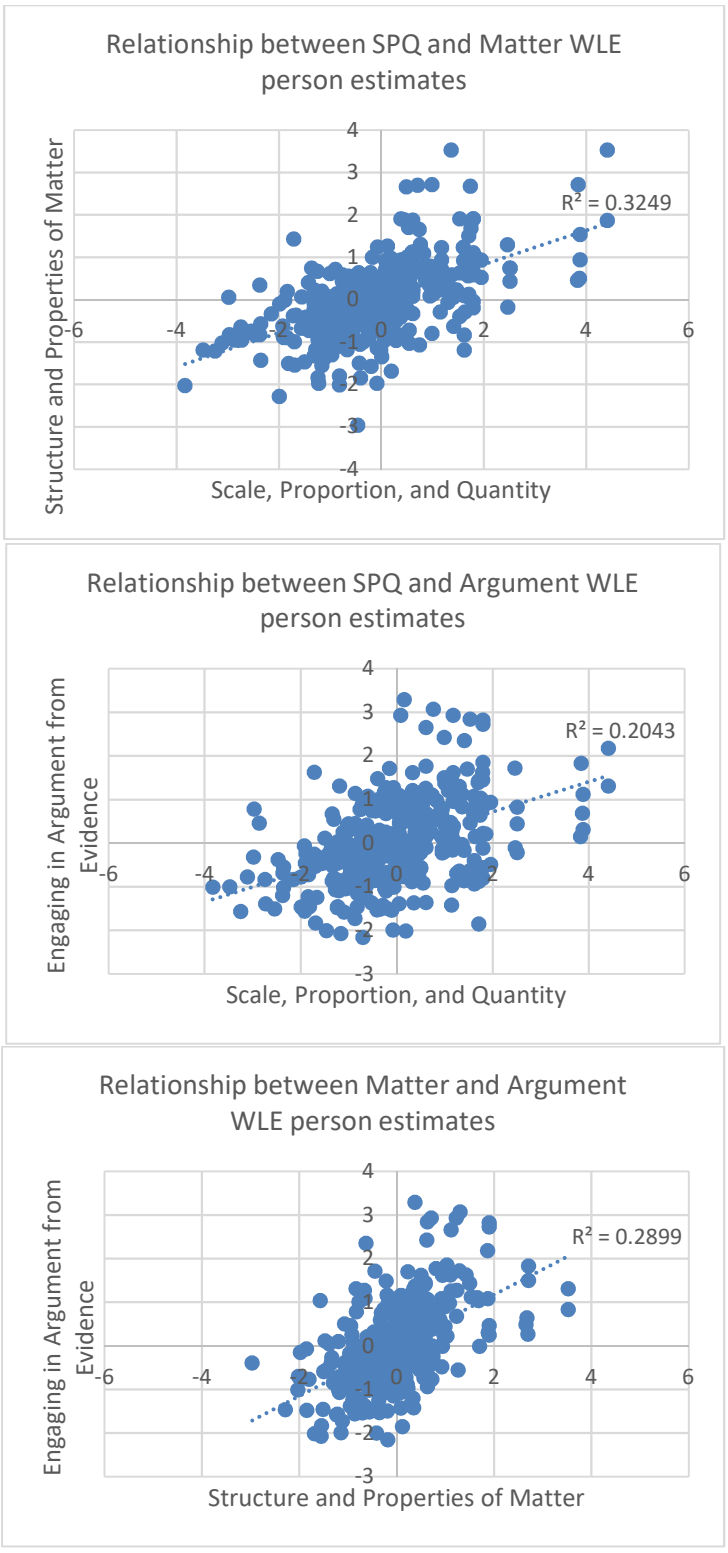


Figure 4.16. Relationship between WLE person estimates for each pair of dimensions.

difference = 72.30, df = 2, $p < 0.01$), and the difference in deviance between the unidimensional and two-dimensional models is much larger than the difference between the two- and three-dimensional models. Thus, the separation of the Engaging in Argument from Evidence dimension has the largest impact on model fit. The impact of establishing separate dimensions for Scale, Proportion, and Quantity, and Structure and Properties of Matter is much smaller.

Person fit. Person fit statistics were reported with the WLE person estimates. ConQuest produces mean square statistic for person fit. Cases with values outside of the range 0.70 through 1.30 were flagged as misfitting, and the percentage of misfitting persons was compared across the three models (Table 4.22). Values lower than 0.70 overfit the model, indicating that the student's responses are more predictable than the model expects, while values higher than 1.30 indicate that the student's responses are less predictable than the model expects (Wright & Linacre, 1994).

The unidimensional model with holistic data was the worst fitting, with the greatest frequency of both overfitting and underfitting persons. The unidimensional model with the multidimensional/analytic dataset had the lowest amount of total misfit. However, the multidimensional model had the lowest number of students with mean squares above 1.30, indicating that the multidimensional model was able to explain more

Table 4.22.

Percentage of Misfitting Persons

	Total Misfit	≤0.70	≥1.30
Analytic Unidimensional	31.98%	23.31%	10.03%
Analytic Multidimensional	41.46%	36.31%	5.42%
Holistic Unidimensional	55.28%	41.73%	13.82%

response variation among students who were considered 'less predictable' in the unidimensional model. These 'less predictable' students are likely students with dissimilar performance on separate dimensions. In addition to providing more nuanced information about these students' performance, the three-dimensional model also improves the fit of their ability estimates.

Heteroscedasticity/homoscedasticity of covariance between dimensions. Scatterplots were created to represent the relationships between each pair of dimensions, based on WLE ability estimates from the multidimensional model (Figure 4.16). The relationship between each pair of dimensions appears to be mostly homoscedastic. There are hints that the covariance may increase at the low and high ends of each scale, but the sparseness of data in those regions of the plots makes it difficult to tell whether this is a meaningful pattern or an illusion. For the most part, the assumption of homoscedasticity seems tenable, meaning that the estimates of covariance/ correlation between the three dimensions may be reasonably assumed to be unbiased and estimates from the multidimensional model are valid.

Variation in the correlation between dimensions for students of different abilities. The dataset was divided into low, medium, and high ability students and the analysis was rerun on each subgroup. The ability groupings and subsequent analysis were repeated for each dimension (i.e., once with ability groups based on ability estimates from the Scale, Proportion, and Quantity dimension, then with ability groups based on the Structure and Properties of Matter dimension, and finally with ability groups based on the Engaging in Argument from Evidence dimension)[32]. Depending on the dimension used as

[32] These smaller subgroup analyses had smaller sample sizes, and therefore all estimates may not be as stable as in the overall analysis.

Table 4.23.

Correlations Among Dimensional Ability Estimates Among High, Medium, and Low Ability Students on Each Dimensions

		Groups divided by ability on SPQ dimension		Groups divided by ability on Matter dimension		Groups divided by ability on Argument dimension	
		SPQ	Matter	SPQ	Matter	SPQ	Matter
Low	Matter	0.51	--	0.90	--	0.64	--
	Argument	-0.24	0.64	0.05	-0.29	-0.03	0.53
Medium	Matter	0.25	--	0.12	--	0.72	--
	Argument	-0.54	0.63	-0.03	0.94	0.21	0.55
High	Matter	0.74	--	0.89	--	0.99	--
	Argument	0.93	0.86	0.77	0.94	0.68	0.61

the basis for grouping by ability, the patterns in correlations among dimensions vary widely (Table 4.23). This contradicts the observed homoscedasticity of the previous scatterplots and suggests that the relationship among the three dimensions is, in fact, non-uniform for students of different abilities. Looking at each pair of dimensions separately reveals some interesting patterns. For the Scale, Proportion, and Quantity, and Engaging in Argument from Evidence dimensions, the correlation between these estimates is weak, or even negative among students with low or medium ability, no matter which dimension is used to define the ability groups. However, among high ability students, the relationship between these dimensions is positive and strong. Looking at the correlations between Scale, Proportion, and Quantity, and Structure and Properties of Matter estimates, the relationship tends to be stronger among the low and high ability groups, but is weaker among students of medium ability. The relationship between Structure and Properties of Matter and Engaging in Argument from Evidence is the most stable. Across all ability groups, the correlations are moderate to high, ranging from 0.52 to 0.94. This

holds regardless of which dimension is used to define the groups, *except* for the low ability group when the matter dimension is used to define the ability groups: in this case, there is a negative correlation between the two sets of estimates. Among students will low ability on the Matter dimension, the relationship between these dimensions may be weak; however, the relationship remains strong among all other students. Overall, the general trend is that strong correlations between dimensions are apparent among students with high ability, but correlations tend to fluctuate or weaken among students with ability classified as "low" or "medium." Potential explanations for this pattern are discussed in Chapter 5.

Reliability. Scale reliabilities for all models are presented in Table 4.24. When a 1-dimensional model is used to scale student responses that were scored with a multidimensional rubric, both WLE and EAP/PV reliabilities are high. When the data is scaled multidimensionally, both WLE and EAP reliability estimates decrease relative to the unidimensional model. The decrement in reliability was much sharper for the WLE person separation estimate. In the most extreme case, WLE reliability on the Scale, Proportion, and Quantity dimension is only 0.38, compared to a reliability of 0.82 when

Table 4.24.

EAP and WLE Person-Separation Reliability for Unidimensional Scales and Multidimensional Subscales with Holistic and Analytic Data

	WLE person separation reliability	EAP/PV reliability
Holistic scoring, 1-dimensional scale	0.79	0.85
Analytic scoring, 1-dimensional scale	0.82	0.86
Analytic scoring, 3 subscales		
Scale, Proportion, and Quantity	0.38	0.74
Structure and Properties of Matter	0.58	0.79
Engaging in Argument from Evidence	0.70	0.78

the items are scaled unidimensionally – a gap of 0.44. EAP reliability also decreases, but the decrement is much smaller (the largest drop is 0.13 on Scale, Proportion, and Quantity). This finding is consistent with other studies, which have found that WLE estimates tend to have lower reliability than EAP estimates (Wang, 2015).

When a holistic scoring rubric and unidimensional model are used, scale reliability is only slightly lower than a unidimensional model with an analytic scoring rubric, and higher than the reliability of individual subscales under the multidimensional approach. This is somewhat surprising, given that the unidimensional/holistic model has many fewer item thresholds than the unidimensional/analytic model, yet the decrement in reliability is fairly small. In fact, the unidimensional/holistic model has a similar number of thresholds to the Engaging in Argument from Evidence subscale, yet the unidimensional/holistic model has a clear advantage in reliability. Although the multidimensional/analytic rubric provides three times the observations (in the form of scores), the corresponding increase in reliability is very small.

Precision of model parameter estimates. When the multidimensional dataset is used, standard errors of the item, rater, and step parameters, and applicable interaction parameters are very similar across the unidimensional and multidimensional models. The multidimensional model has slightly smaller standard errors across the board, but the differences are negligible. Therefore, neither model demonstrates a strong advantage with regard to the precision of parameter estimates.

When the holistic dataset is used, standard errors of the parameter estimates are similar in size to those when the multidimensional dataset is used. Regardless of the chosen model/scoring rubric, the precision of item location estimates is similar.

Precision of person ability estimates. EAP and WLE person ability estimates from the unidimensional model tended to have smaller standard errors of measurement than EAP and WLE estimates from the subscales of the multidimensional model (Figures 4.17 and 4.18). The size of the errors varied by subscale. Scale, Proportion, and Quantity standard errors tended to be much larger (average EAP SEM = 0.52; WLE = 0.92), while the Structure and Properties of Matter (average EAP SEM = 0.34; WLE = 0.53) and Engaging in Argument from Evidence errors (average EAP SEM = 0.37; WLE = 0.49) were a bit smaller. This difference is in line with expectation, as scales with a smaller number of items tend to have a larger standard error of measurement (Briggs & Wilson, 2003). The unidimensional scale had the largest number of item thresholds/score points and the lowest standard errors (average EAP SEM = 0.28; WLE = 0.32). Of the three subscales, the Scale, Proportion, and Quantity dimension had the smallest number of items and the highest standard errors. This indicates that EAP and WLE person estimates from the unidimensional scale are more precise than the corresponding estimates from the multidimensional subscales. The loss in precision from the multidimensional estimates is most extreme on the Scale, Proportion, and Quantity dimension.

When the holistic scoring rubric was used in conjunction with a unidimensional Rasch model, the average EAP standard error was 0.34 and the average WLE standard error was 0.40. Unlike in the reliability analysis, the unidimensional model/ multidimensional dataset combination holds a clear advantage over the holistic dataset. The difference between the two models is probably due at least in part to the larger number of items in the multidimensional dataset. However, it should be noted that the holistic model standard errors are similar in size to those from the Structure and

Figure 4.17. Scatterplots of EAP estimates and standard errors for all scales.

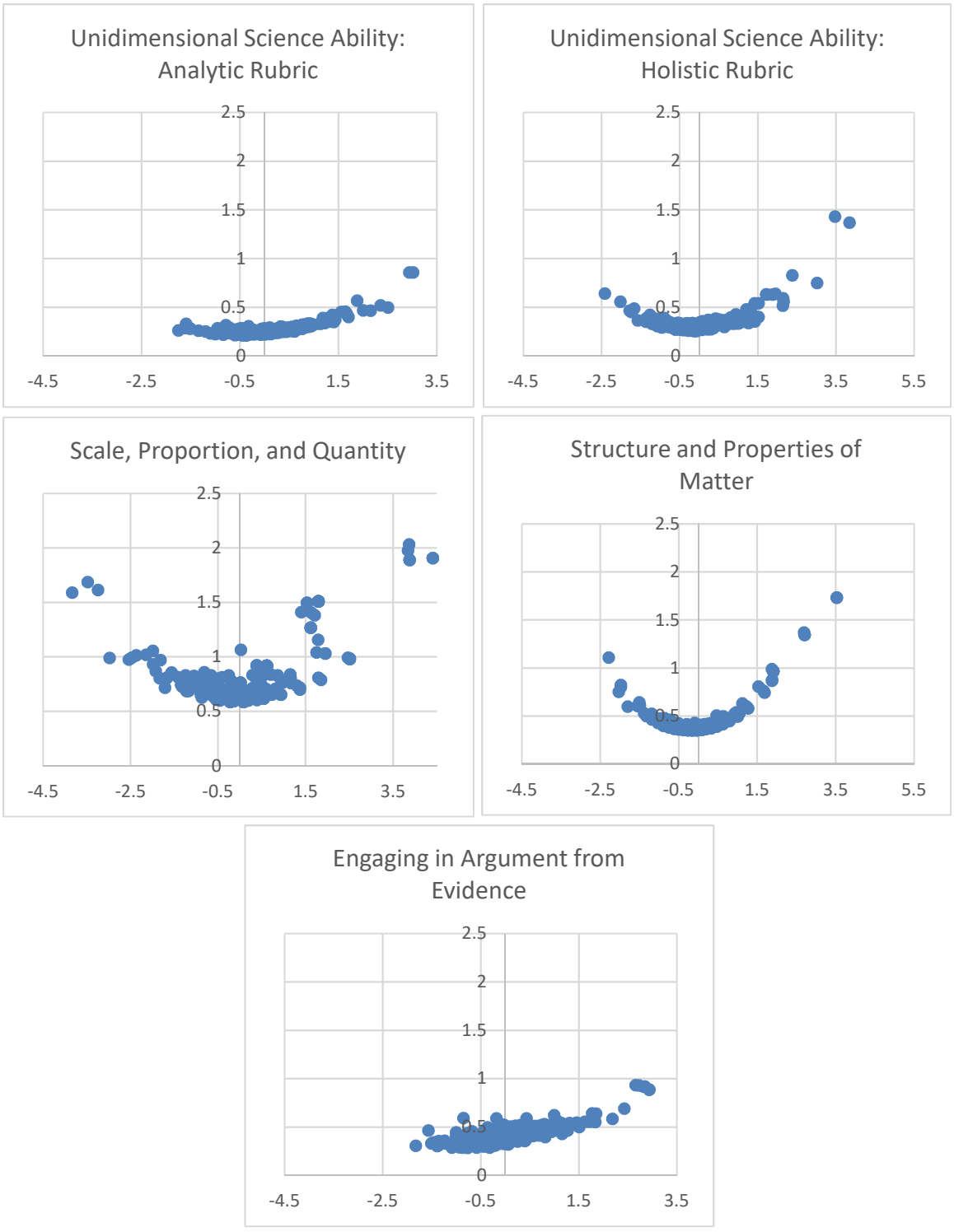


Figure 4.18. Scatterplots of WLE estimates and standard errors for all scales.

Properties of Matter and Engaging in Argument from Evidence subscales, which have a similar number of items (except for WLE standard errors, which are higher among the subscales). Therefore, it appears that the individual subscale estimates are similar in precision to the holistic estimates (i.e., provide just as much information, measured by the information function which is the inverse of the standard error), even though they ostensibly evaluate only one aspect of the student's responses where the holistic estimates evaluate the entire response. In terms of information, the multidimensional rubric is better than the holistic rubric when a unidimensional model structure is used. Arguably, the multidimensional rubric is better than the holistic rubric, even when a multidimensional model structure is used, since it provides information with a similar level of precision, but about multiple subscales.

Research question 4: How well does student performance data reflect the hypothesized underlying construct framework?

Item difficulty estimates. Depending on the chosen dimensionality structure, there are differences in the concordance of the range of person ability estimates and item difficulty estimates. Ideally, the distribution of item difficulty will mirror the distribution of person ability because items are most informative when they are close in difficulty to the examinee's ability. When a unidimensional model is used, the distribution of person ability and item difficulty estimates appears close to ideal. The distribution of person ability is approximately normal, with most of the examinees falling near the scale average of zero (Figure 4.19). The distribution of item difficulty estimates mirrors the person ability distribution. The majority of item thresholds fall near the scale average, with ample numbers of thresholds extending to the upper and lower extremes.

When the items and person ability estimates are separated into three related scales from a multidimensional model, the concordance between item difficulty and person ability lessens somewhat, indicating that the unidimensional structure actually masks some gaps in item coverage, which are revealed when the items are separated into subscales. These gaps are particularly prevalent on the Structure and Properties of Matter and Engaging in Argument from Evidence Dimensions (Figure 4.20). The Scale, Proportion, and Quantity scale items are appropriately distributed along the scale, except for a gap at the bottom of the scale below -2.50 logits. The Structure and Properties of Matter dimension had a condensed range of item difficulty estimates. Most of the item thresholds are clustered near the middle of the scale, with only two thresholds larger than 0.7 logits and none lower than -1.10 logits. Many students demonstrated ability lower than the Matter items were able to measure well, and there were very few items that targeted students with high ability. Finally, on the Engaging in Argument from Evidence dimension, most of the item thresholds clustered near the bottom of the scale, lower than most students' estimated ability. There were fewer thresholds aligned with the bulk of the ability distribution, and very few thresholds above 1 logit, near the higher-ability students.

Without examining the separate subscale distributions, it would appear that items are well-matched to student performance on the overall construct, when in actuality there are few items able to differentiate between degrees of sophistication in students' understanding of fundamental matter concepts or argumentation ability at the high and low ends of the scale. The multidimensional model thus provides important feedback for

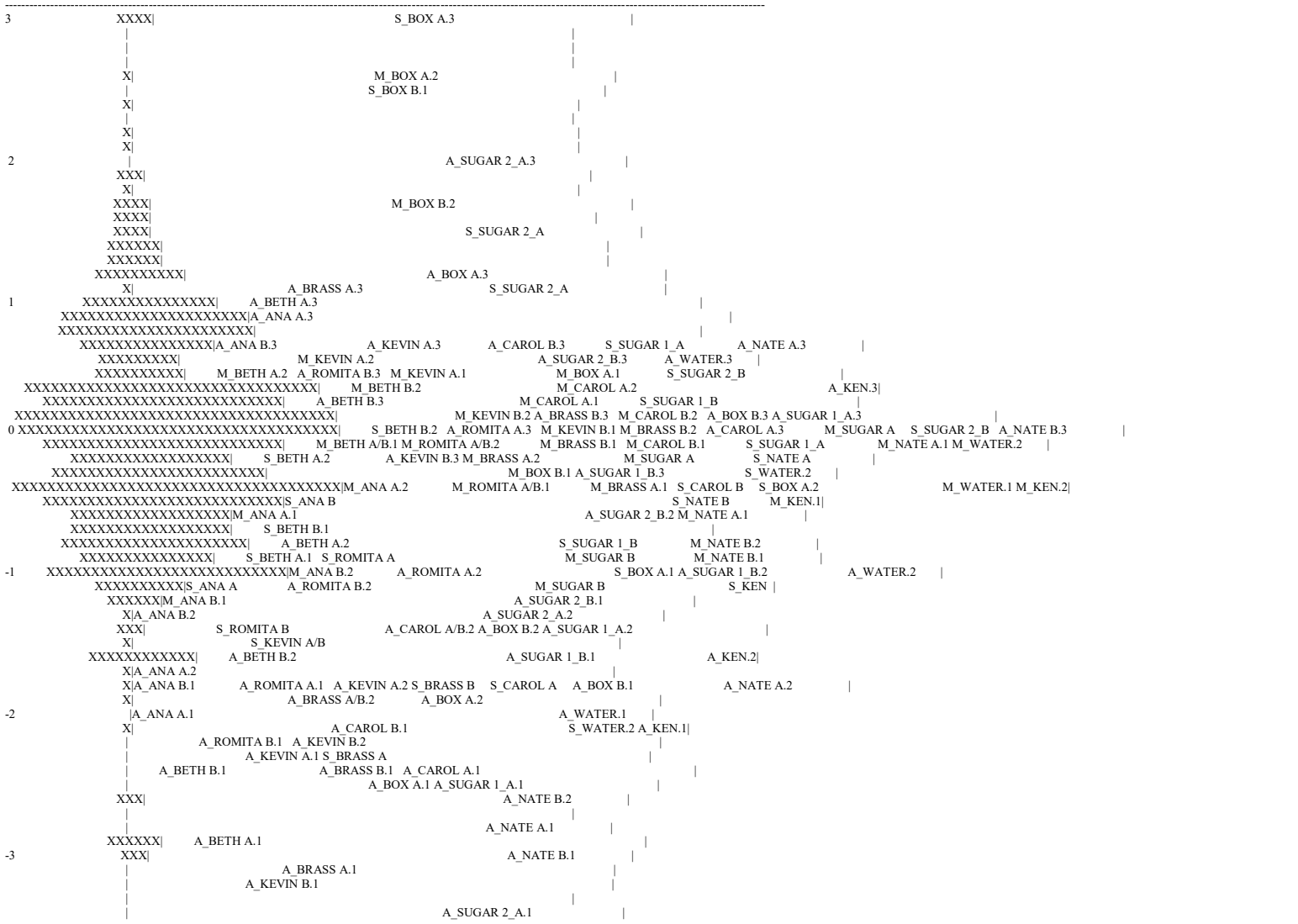


Figure 4.19. Wright Map from the unidimensional model and multidimensional scoring rubric

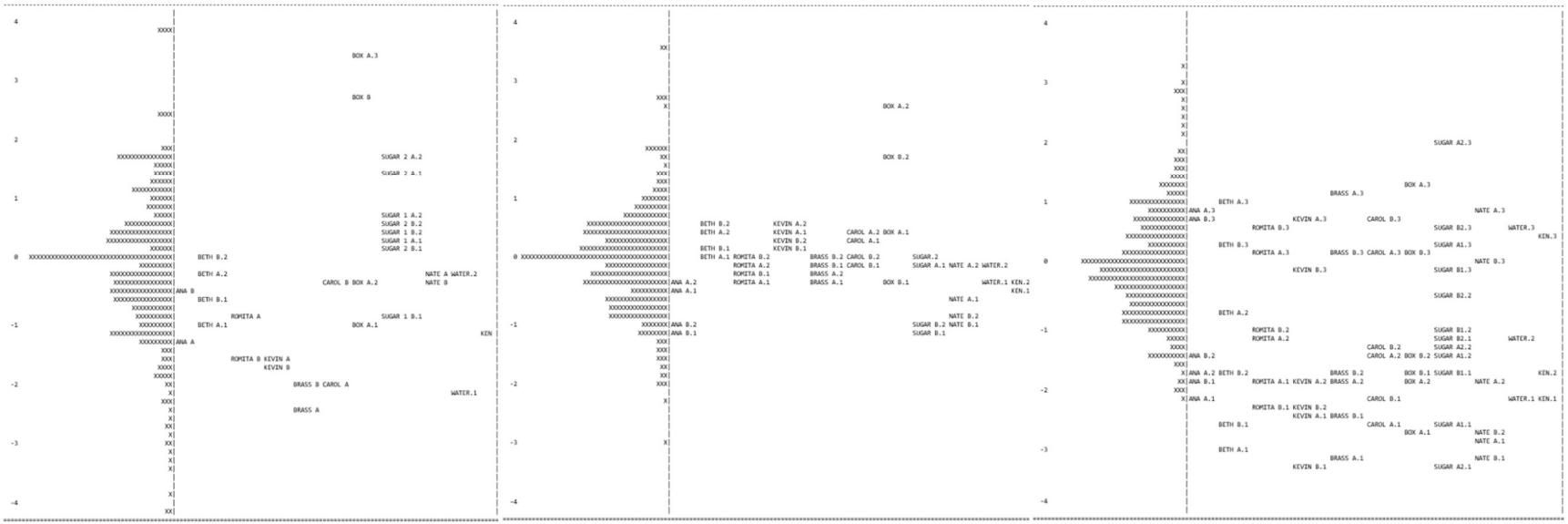


Figure 4.20. Distribution of items (right sub-panel) and persons (left sub-panel) for the three assessment subdimensions: Scale, Proportion, and Quantity, Structure and Properties of Matter, and Engaging in Argument from Evidence, respectively, from left to right.

assessment of multiple related constructs, providing suggestions for areas for improvement of item coverage that would not be evident under a unidimensional model.

Person ability estimates. When the multidimensional dataset is used in conjunction with a unidimensional model, the distribution of person ability is approximately normal and ranges from about -3 logits to +3 logits (Figure 4.19).

Splitting the scale into three sub-dimensions has varying effects, depending on the specific subscale (Figure 4.20). On the Scale, Proportion, and Quantity dimension the person ability distribution is flattened, with student estimates distributed across a broader logit range. This suggests that the dimension measures a sub-construct with significant, measurable variability, and may be worth separating from the other dimensions. The Structure and Properties of Matter and Engaging in Argument from Evidence distributions both have larger peaks and shorter tails than the Scale, Proportion, and Quantity dimension. The Engaging in Argument from Evidence distribution, in particular, has an almost nonexistent lower tail, with no student estimates below -2 logits. These gaps in the student ability distribution may reflect poor construct definition and/or item design; in particular, that the items do not differentiate well between students who fall in those ranges of the subscale ability distribution.

Order of item difficulty estimates. In Figures 4.21 and 4.22, each item is placed on the dimensional subscales in its hypothesized location according to the relative difficulty of the concept(s) assessed by that item, determined by the hypothesized construct definition presented in Chapter 1. The text of all assessment items can be found in Appendix F.

4	Weight is invariant during phase change (freezing, melting, evaporating, condensation) **(CAROL)**.
3	Integrates weight, volume, and heaviness for size in compositional model of materials **(ROMITA, BRASS, BOX)**. Knows any solid, liquid, or granular sample, however small, has weight and volume **(SUGAR, WATER)**. Knows tiny things have weight, and weight is invariant across crushing and dissolving **(KEVIN)**. Differentiates volume from area and understands volume of solid objects and liquids is invariant with reshaping **(NATE)**. Understands water displacement depends upon volume of submerged object **(BETH)**.
1	Has initial concept of amount of material as a quantity that remains invariant when object changes appearance with reshaping, because nothing is added/removed. **(ANA, KEN)**

Figure 4.21. Hypothesized location of items: Structure and Properties of Matter dimension. Items are listed next to the relevant concept. All items within a particular level are hypothesized to have roughly the same difficulty. Note that non-assessed concepts were not included in this figure. For a complete list of all concepts included in the Structure and Properties of Matter construct, please see Chapter 1.

On the Engaging in Argument from Evidence dimension, there is no hypothesized progression of items, but instead a hypothesized progression of scoring categories. It is hypothesized that it will be easier to provide a claim without evidence or with weak evidence/poor reasoning (Threshold 1) across all items, and most difficult to provide an argument that includes both evidence and reasoning (Threshold 3) across all items.

Structure and Properties of Matter. Figure 4.23 shows the observed difficulties for all items on the Structure and Properties of Matter scale. Nearly all scale items were hypothesized to have similar difficulty, as they all measured concepts contained in Level 3 of the construct map. In line with this prediction, the range of item estimates was fairly compressed, with most items falling between -1 and +1 logits. The three items that were hypothesized to be the easiest did fall near the bottom of the scale. However, the two items that were hypothesized to be the hardest ended up near the middle of the scale. These two items (CAROL A and CAROL B) assess whether students recognize that

4	Use addition and subtraction to mathematically reason about volume (e.g., infer volume from the change in water level **(NATE)**, use water displacement to measure the volume of solid, liquid, and granular materials **(BETH)**). Uses multiplication and division to mathematically reason using measurements (e.g., infer the weight and volume of proportionally smaller/lighter or larger/heavier objects made of the same material.) **(BOX A.3, BOX B, WATER.2, SUGAR 1.2, SUGAR 2.2)**
3	Uses addition and subtraction to mathematically reason about weight (e.g., infer the weight of liquids in a container). **(KEVIN, CAROL)** Understands the structure of 3D arrays, and can use centimeter cubes to measure volume by comparison. **(KEN, ANA)** Has generalized knowledge of fractions (there are an infinite number of fractions between any two integers); number and measure line is a dense, and quantities form a continuum **(SUGAR 1.1, SUGAR 2.1)** Beginning understanding of measured characteristics that involve proportional relationships (i.e., can use relative scales to describe speed and density when distance/weight or time/volume is held constant.) **(BOX A.2)**
2	Knows how to measure weight with balance scale, length with rulers, and area with tiles. **(BRASS)** Understands measures using scales, rulers are more reliable than senses. **(ROMITA)** Understands the structure of 2D arrays, and can use centimeter cubes to measure area. Has knowledge of a few special fractions (1/2, 1/4); number and measure line is spotty. **(WATER.1)** Can use measures to evaluate relative scale (for weight, length, area). **(BOX A.1)**

Figure 4.22. Hypothesized location of items: Scale, Proportion, and Quantity dimension. Items are listed next to the relevant concept. All items within a particular level are hypothesized to have roughly the same difficulty. Note that non-assessed concepts were not included in this figure. For a complete list of all concepts included in the Scale, Proportion, and Quantity construct, please see Chapter 1.

weight of a material remains the same after melting. Because this concept involves a phase change, it was hypothesized to be more difficult than two similar items (KEVIN A and KEVIN B) which tested whether students recognized that weight was invariant when a solid material was crushed. In fact, it was no more difficult. Recognizing the invariance of weight during a phase change may not be as difficult as previously thought.

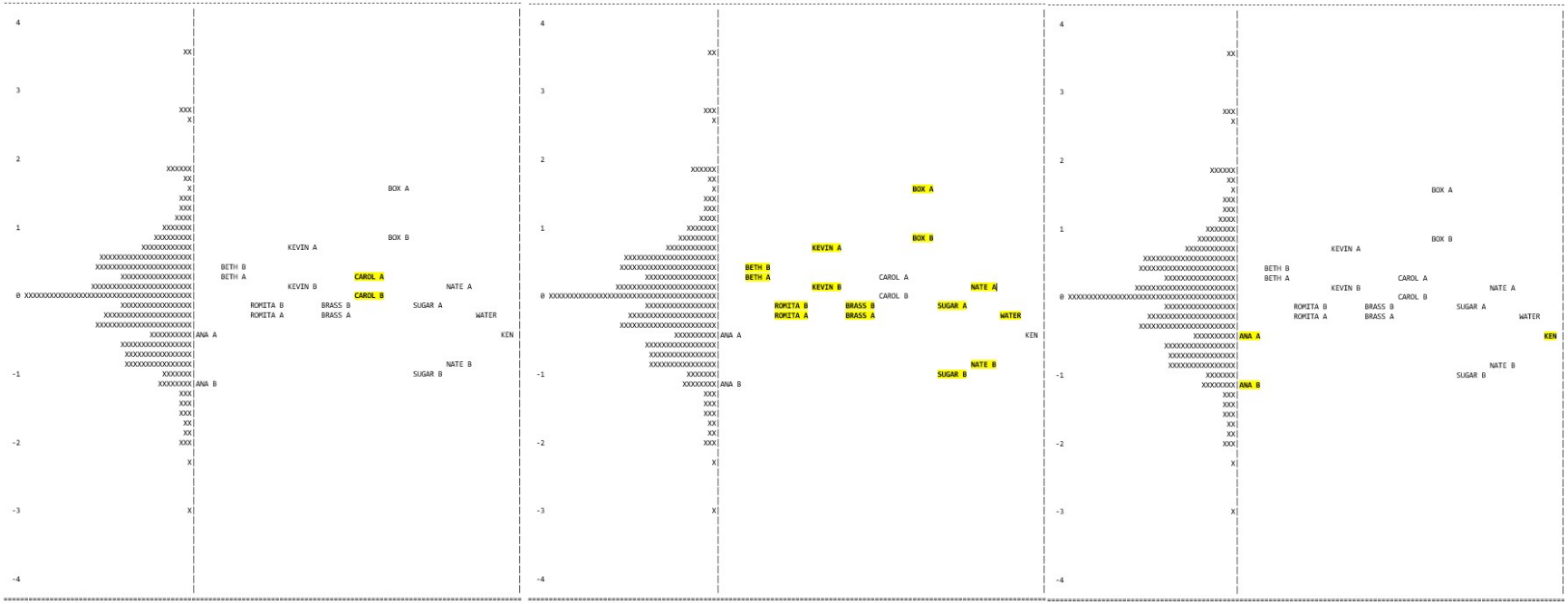


Figure 4.23. Actual distribution of item difficulty on the Structure and Properties of Matter dimension. The three panels highlight the locations of the hypothesized high, medium, and low difficulty items, respectively, from left to right.

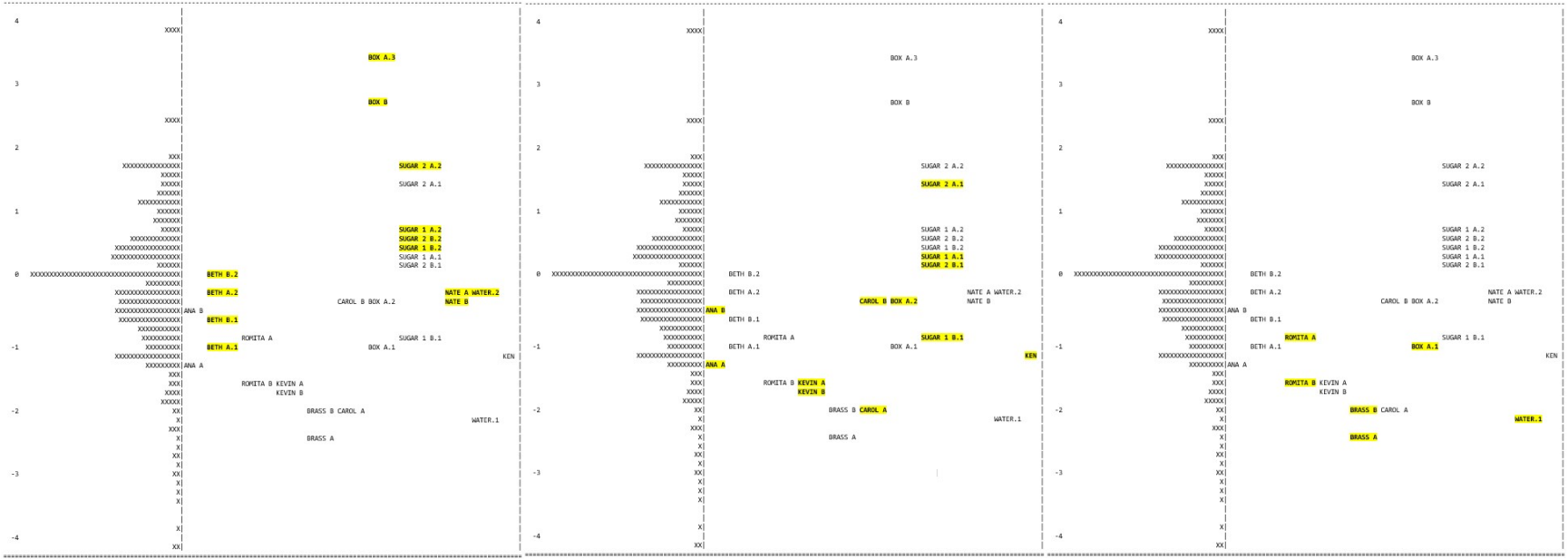


Figure 4.24. Actual distribution of item difficulty on the Scale, Proportion, and Quantity dimension. The three panels highlight the locations of the hypothesized high, medium, and low difficulty thresholds, respectively, from left to right.

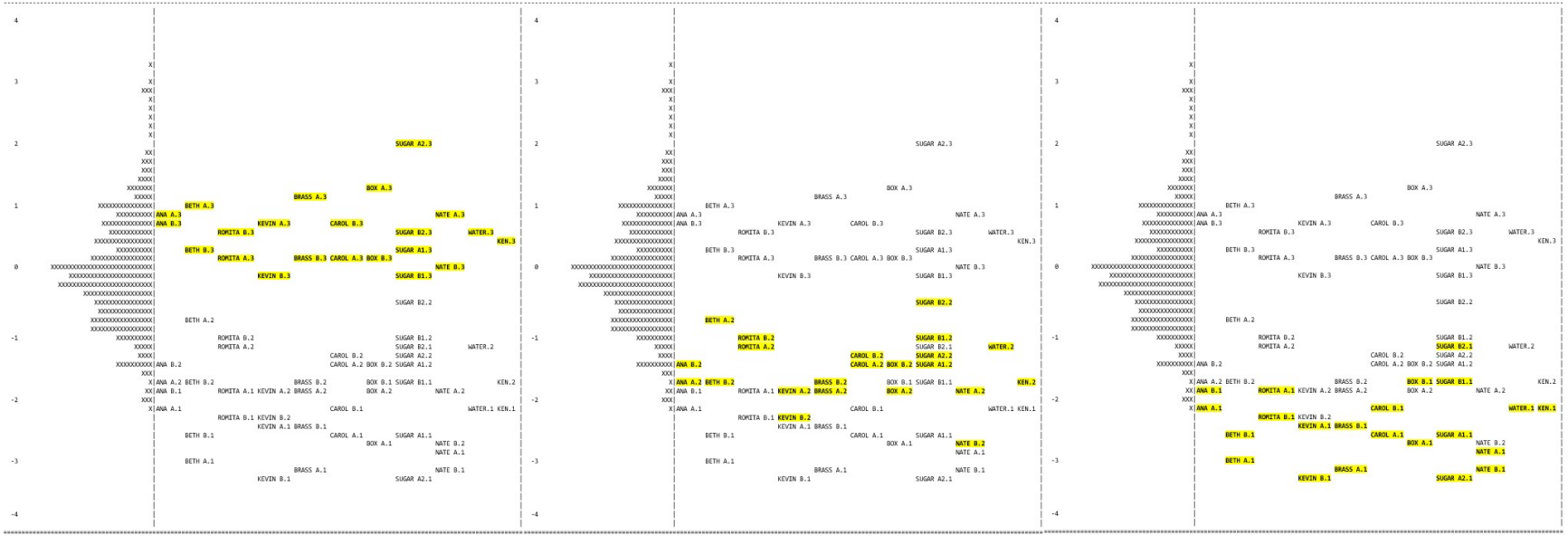


Figure 4.25. Actual distribution of item difficulty on the Engaging in Argument from Evidence dimension. The three panels highlight the locations of the hypothesized high, medium, and low difficulty thresholds, respectively, from left to right.

Other discrepancies from the hypothesized order included BOX A and BOX B, the most difficult items on the assessment. These two items assessed whether or not students could integrate information about volume and weight to make a judgment about whether two objects could be constituted of a common material. This concept was shared with four other items on the assessment, ROMITA A and B, and BRASS A and B, all of which were substantially easier than BOX A and BOX B despite measuring the same concept. The difference is that ROMITA and BRASS both kept one of the two factors (weight or volume) constant, while the objects in BOX had different weights and different volumes. It appears that integrating weight and volume to make judgments about materials may be easier or more difficult depending on the number of varying parameters that students have to consider.

Finally, it must be noted that differences in item scaffolding may have affected the difficulty estimates of certain items, affecting the distribution of item difficulty.

Scale, Proportion, and Quantity. Figure 4.24 shows the observed difficulties for all items on the Scale, Proportion, and Quantity scale. On a broader level, it seems that the hypothesized distribution of item difficulties mostly holds up. All of the items measuring Level 4 SPQ concepts tend to be located near the top of the scale, items measuring Level 3 concepts tend to be located mid-scale, and items measuring Level 2 concepts tend to be located near the bottom. There are some inconsistencies between the hypothesized and actual locations. For example, some of the lowest SUGAR thresholds appear near the top of the scale, despite being hypothesized to be mid-scale items. Others, like BETH and NATE which were hypothesized to be more difficult, appeared closer to the middle of the scale. These discrepancies have two potential explanations. One is that

item scaffolding affected the observed distribution of item difficulty, but the previous analysis of the association between scaffolding and item difficulty did not provide clear evidence in support of this explanation. The other potential explanation is that differences in the item tasks contributed to the dispersion of item difficulty estimates. For example, of the items measuring Level 4 concepts, the two easiest items required students to calculate changes in volume using subtraction. The most difficult items asked students to use division and proportions. Perhaps the increasing complexity of the mathematical calculation required for the task is a bigger contributor to task difficulty than was reflected in the hypothesized progression. Although the hypothesized construct map was largely supported by the actual distribution of item difficulty, a few discrepancies suggest areas for potential revision of the construct definition.

Engaging in Argument from Evidence. Figure 4.25 shows the observed difficulties for all items on the Engaging in Argument from Evidence scale. The thresholds are clearly separated into three regions. For the most part, the group of first thresholds falls below the group of second thresholds, which falls below the group of third thresholds. There is some overlap between the ranges of threshold groups 1 and 2, and the distance between the first and second thresholds for a particular item tend to be smaller than the distance between the second and third thresholds for that item. This means that the leap from presenting a weak argument to presenting an argument based on one valid component (either evidence or reasoning) seems less difficult than the leap to presenting an argument based on two valid components (both evidence and reasoning). Within each threshold, there are a range of difficulties, suggesting that there are some

item-specific factors that cause variation in the difficulty of reaching each threshold. Overall, the observed item difficulties support the hypothesized construct definition.

Item fit. For reference, all items included in the pilot assessment can be found in Appendix F. The mean square statistic provides an indication of how well individual items contribute to overall model fit by looking at differences between expected student performance (based on item difficulty and student ability) and observed student performance. Of the 64 scale items, 14 items were flagged as having significant misfit. One of these (KEVIN A) was explored in detail in an earlier section; it is likely that the open-ended response format led to confusion since an alternate version of the item with a selected-response format fit the model well. For the remaining 13 flagged items, item wording and content were examined to make a judgment about whether or not the items are appropriate for continued use in assessment.

BRASS A (Figure 4.26) was flagged as misfitting on the Structure and Properties of Matter dimension, and its counterpart (BRASS B) demonstrated similar misfit statistics despite falling just below the threshold for flagging. Based on examination of the item, there is no obvious violation of best testing practices. The wording and images are clear, content-specific terms are defined, and all information needed to answer the question is provided in the stimulus material. It is possible that the distracters are the source of misfit; the third and fourth answer options, in particular, reflect popular misconceptions, and they may be attractive distracters. However, students' selection of these answers reflect an important gap in their understanding, and an item should not be removed from assessment just for having an attractive distracter, so long as correct and incorrect answers are valid indicators of the student's proficiency with the underlying

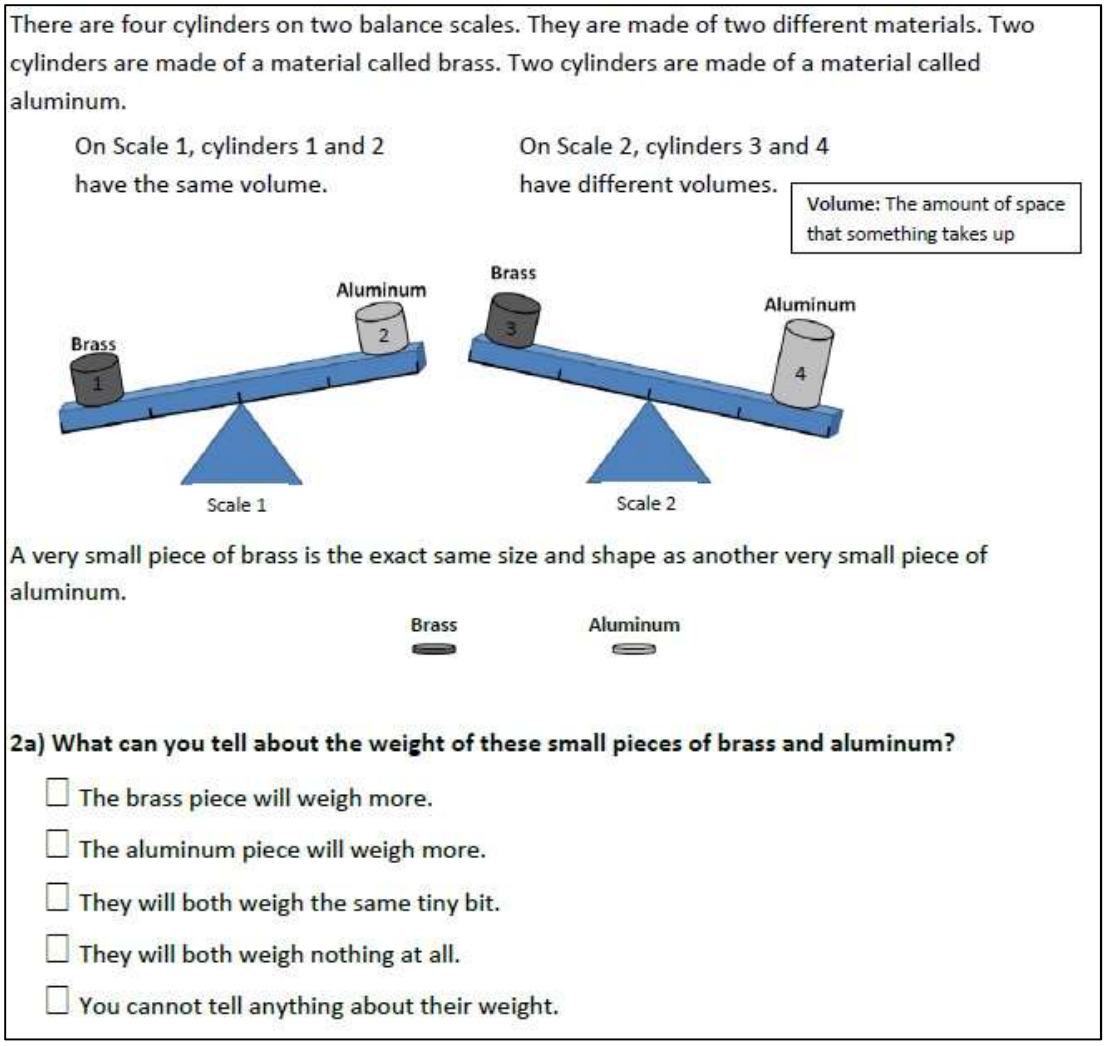


Figure 4.26. The Structure and Properties of Matter sub-prompt of BRASS A, an item flagged for misfit.

concept. With no clear source of information about the item's failings, curriculum and content experts should be consulted before making a decision about retaining the item.

On the Engaging in Argument from Evidence dimension, a majority of items (12 of 22) were flagged as misfitting. Previous analysis (RQ 3) focused on the impact of scaffolding and response format, and demonstrated that selected-response arguments tend to have larger fit statistics than constructed-response items. However, it is worth noting that several constructed-response items were also flagged as misfitting, indicating that

response format is not the only source of misfit. In the previous analysis for research question 1 and 2, low interrater reliability was observed on the Engaging in Argument from Evidence dimension. Problems in the argument rubric might also lead to misfit, as well as unreliability. The raters' judgments, if based on extraneous factors unrelated to students' underlying ability, would be more difficult to predict, therefore leading to larger discrepancies between the expected and observed responses and inflating item fit statistics. Reexamination and revision of the scoring rubrics and construct definition for the Engaging in Argument from Evidence dimension are necessary; until then, item fit statistics reveal more about the overall quality of the Engaging in Argument from Evidence scale than about problems with individual items.

Standard error of person ability estimates. On the unidimensional scale, and the Structure and Properties of Matter and Scale, Proportion, and Quantity subscales, standard errors tended to be higher for examinees at the extreme ends of the scale, as expected (Wilson, 2005) (see Figures 4.17 and 4.18). On the Engaging in Argument from Evidence subscale, this pattern was muted, suggesting that the students' locations are not related to the amount of error in their ability estimates. This is likely because of the large number of items at the bottom of the Engaging in Argument from Evidence subscale. Because there were so many items, there was more information available about low-performing Engaging in Argument from Evidence students, mitigating the typical increase in standard error for students at the extreme ends of the scale.

Differential Item Functioning. DIF was examined by fitting a facet model with ConQuest Version 4 (Adams, Wu, & Wilson, 2015). The subgroup characteristic of

interest was added as an additional predictor in the item response model, and differences in item difficulty between different subgroups were directly compared.

Grade level. Output from the DIF analysis for grade and Inquiry Project participation may be found in Appendix E. On the Structure and Properties of Matter dimension, there is a 0.23 logit difference in ability for students in the 4th and 5th grade, such that 5th graders tend to have better performance, on average. This difference is of moderate size: almost a third of a standard deviation of the Structure and Properties of Matter dimension (subscale variance was 0.59). This difference does not indicate DIF, but only confirms the intuitive prediction that 5th graders would perform better than 4th graders on the test.

Based on the overall item difficulty parameters, 18 of the 20 items do not demonstrate a significant difference in item difficulty between grades. There are two items where the difference between difficulty estimates for 4th and 5th grades are statistically significantly different: item 5B and item 9A. Item 5B was easier for 4th graders and item 9A was easier for 5th graders. For both of these items, an alternate version of the item was also tested with an identical matter prompt in both framing and response format. In both cases, the alternate version did not demonstrate DIF, suggesting that the differences observed here may be due to chance.

Furthermore, there does not appear to be any difference in the step structure for 4th and 5th graders, as the difference in model deviance between a model with a step structure that varies for 4th and 5th grades and a model with an invariant step structure across grades is not statistically significant (χ^2 difference = 22.48, df = 20, p = 0.32).

On the Scale, Proportion, and Quantity dimension, there is a 0.24 logit difference in ability for students in the 4th and 5th grade, such that 5th graders tend to have better performance, on average. This difference is of moderate size: about a quarter of a standard deviation of the Scale, Proportion, and Quantity dimension (subscale variance was 1.11). This difference does not indicate DIF, but again confirms that 5th graders performed better than 4th graders on this dimension.

Based on the overall item difficulty parameters, there are two items where the difference between difficulty estimates for 4th and 5th grades are statistically significantly different: item 2A and item 7B. Item 2A was easier for 4th graders. For this item, an alternate version was tested with an identical Scale, Proportion, and Quantity prompt in both framing and response format, and the alternate version did not demonstrate DIF. This suggests that the difference observed here may be due to chance. Item 7B was easier for 5th graders, and the difference between grade estimates was 1.83 logits. If all items demonstrated a difference of that magnitude between grades, it would shift the student distribution by about 1.75 standard deviations – a very large amount. In this case, an alternate version of the item used a different prompt and response format, so this item may provide a legitimate source of DIF. The simplest explanation is that the item targets a learned concept; item 7B asks students to consider whether several objects of various weights and volumes may be made of the same material. This is a very difficult item, which requires an understanding of both density and proportionality in order to respond correctly. If students have not been exposed to one or both of the concepts, it may make the item more difficult. The *Inquiry Project* curriculum introduces the idea of density, or "heaviness for size", in the 4th grade, and fractions/proportions are often introduced in

mathematics classes around the same time. The pilot test was administered at the beginning of the school year, meaning that 4th grade students in the sample may not have been exposed to these concepts yet. This may have affected their ability to respond to the item, and would explain the observed difference in difficulty.

Finally, there does not appear to be any difference in the step structure for 4th and 5th graders, as the difference in model deviance between a model with a step structure that varies for 4th and 5th grades and a model with an invariant step structure across grades is not statistically significant (χ^2 difference = 7.75, df = 30, p = 0.99).

On the Engaging in Argument from Evidence dimension, there is a 0.30 logit difference in ability for students in the 4th and 5th grade, such that 5th graders tend to have better performance, on average. This difference is statistically significant and of moderate size: almost two-fifths of a standard deviation of the Engaging in Argument from Evidence dimension (subscale variance was previously reported at 0.67). This difference does not indicate DIF, but only confirms that 5th graders had better performance than 4th graders on this dimension.

Based on the overall item difficulty parameters, there are six items where the difference between difficulty estimates for 4th and 5th grades are statistically significantly different. Items 3B, 4B, and 5B are easier for 4th grade students, whereas items 6B, 10, and 11 are more difficult for 4th grade students on the Argument dimension. It is unclear why these items demonstrate this pattern. The observed DIF may have something to do with response format; 2 of the 3 items that were easier for 4th graders employed a multiple choice format, whereas all 3 of the items that were easier for 5th graders employed an open response format. However, several other items utilized these response

formats but were not flagged for DIF, leaving the exact source of the observed DIF unclear. Furthermore, given that there are unresolved issues with the Engaging in Argument from Evidence scale, namely, the lack of interrater reliability, it makes sense to reexamine DIF once the noise in scoring has been reduced.

There does not appear to be any difference in the step structure for 4th and 5th graders, as the difference in model deviance between a model with a step structure that varies for 4th and 5th grades and a model with an invariant step structure across grades is not statistically significant for $\alpha = 0.05$ (χ^2 difference = 57.98, df = 44, p = 0.08).

Inquiry Project participation. On the Structure and Properties of Matter dimension, there is a 0.03 logit difference in average ability between students who participated in the Inquiry Project and students who did not. This difference is very small, and is not statistically significant. However, the composition of the two groups was not equal – the *Inquiry Project* group included many more 4th graders than the non-Inquiry group, and the *Inquiry* and non-*Inquiry* samples came from different regions with different socioeconomic characteristics. Student background characteristics (e.g., socioeconomic status, previous performance in science) were not accounted for in this comparison.

When the *Inquiry* and non-*Inquiry* samples are limited to 5th grade students only, there is a 0.31 difference, with *Inquiry Project* 5th graders performing better than non-*Inquiry* 5th graders. This difference is of moderate size: about two-fifths of a standard deviation of the Structure and Properties of Matter dimension (subscale variance was previously reported at 0.60).

Based on the overall item difficulty parameters from the grade 5 sample, there are five items where the difference between difficulty estimates for *Inquiry* and non-*Inquiry* students are statistically significantly: items 1A and 1B, 8A, 10, and 11. Items 1A, 1B, and 11 all assessed whether students recognized that a 3D object retained its volume after being rearranged, and all were more difficult for students who had taken the *Inquiry Project* curriculum. One potential explanation for this surprising finding is a failure of curriculum implementation, which is discussed further in Chapter 5. On the other hand, items 8A and 10 both assessed whether students recognized that very tiny objects have both weight and volume, and both items were easier for students who had taken the *Inquiry Project* curriculum. This finding is in line with the expectation that curriculum participation would offer students an advantage on items that assess curriculum-specific concepts.

There does not appear to be any difference in the step structure for *Inquiry* and non-*Inquiry* students, as the difference in model deviance between a model with a step structure that varies for *Inquiry* and non-*Inquiry* students and a model with an invariant step structure across curricula is not statistically significant (χ^2 difference = 9.94, df = 20, p = 0.99).

On the Scale, Proportion, and Quantity dimension, there is a 0.20 logit difference in ability for students who participated in the *Inquiry Project* and students who did not, such that students who took the *Inquiry Project* curriculum performed better than students who did not. This difference is present despite the unequal sample composition; the *Inquiry Project* group was younger, on average, but still demonstrated higher performance on the Scale, Proportion, and Quantity dimension. This difference is

statistically significant and of moderate size: about a fifth of a standard deviation of the Scale, Proportion, and Quantity dimension (subscale variance was previously reported at 1.11). When the analysis is limited to only 5th grade students in each group, the gap between the groups increases substantially to 0.79 logits, or 75% of the scale's standard deviation. This difference is statistically significant. Again, *Inquiry Project* students found the items easier overall than non-*Inquiry* students.

Based on the overall item difficulty parameters from the grade 5 sample, there are two items where the difference between difficulty estimates for *Inquiry* and non-*Inquiry* students are statistically significantly different: item 4B and item 10. Both items were easier for *Inquiry Project* students. Item 4B asks students to calculate the weight of an object by subtracting the weight of its container. Item 10 asks students to calculate the weight of a single drop of water, given the weight of 20 drops. Both of these skills are covered during the *Inquiry Project* curriculum, therefore the most likely explanation for the difference in performance is that the curriculum introduced concepts that allowed students to perform better on these items. However, there may be other explanations for the gap in item difficulty, such as other contextual factors that varied along with curriculum participation (i.e., state of residence).

Finally, there does not appear to be any difference in the step structure for Inquiry and non-Inquiry students, as the difference in model deviance between a model with a step structure that varies based on curriculum and a model with an invariant step structure across curricula is not statistically significant (χ^2 difference = 12.36, df = 9, p = 0.20).

On the Engaging in Argument from Evidence dimension, there is a 0.232 logit difference in ability for students who participated in the *Inquiry Project* and students who

did not, such that students who took the *Inquiry Project* curriculum performed better than students who did not. This difference is present despite the unequal sample composition; the *Inquiry Project* group was younger, on average, but still demonstrated higher performance on the Engaging in Argument from Evidence dimension. This difference is statistically significant and of moderate size: about a third of a standard deviation of the Engaging in Argument from Evidence dimension (subscale variance was 0.67). When the analysis is limited to only 5th grade students in each group, the gap between the groups increases slightly to 0.456 logits, or just over half of the scale's standard deviation. This difference is also statistically significant. Again, *Inquiry Project* students found the items easier than non-*Inquiry* students.

Based on the overall item difficulty parameters from the Grade 5 sample, there are four items where the difference between difficulty estimates for *Inquiry* and non-*Inquiry* students are statistically significantly different: item 4B, 6B, 8B, and 9B. Of these, only 8B is easier for *Inquiry* students; the remainder are more difficult. Items 4B and 9B both ask the students to create an argument by selecting from among given response options. Items 6B and 8B are open-ended single-prompt arguments, in which students are asked to combine all dimensions of their response as part of a single argument. The four item scenarios do not share a common concept. Overall, there does not seem to be a pattern related to response format or a shared concept that might explain the difference in performance on these items, so the source of the DIF remains unclear. Given that there are unresolved issues with the Engaging in Argument from Evidence scale, namely, the lack of interrater reliability, it makes sense to reexamine for patterns in DIF among these items once the noise in scoring has been reduced.

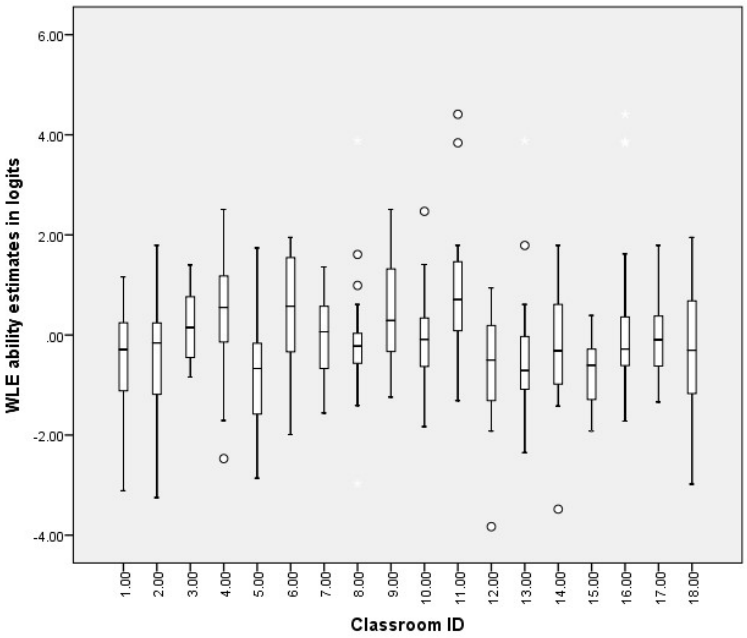


Figure 4.27. Student ability boxplots, grouped by classroom, for the Scale, Proportion, and Quantity dimension. Classrooms 1-10 were *Inquiry Project* participants, while classrooms 11-18 were not. Classrooms 4, 9, 10, 12, and 14-18 are 5th grade classrooms, classrooms 1-3 and 5-8 are 4th grade classrooms, and classrooms 11 and 13 are mixed classrooms.

Finally, there does not appear to be any difference in the step structure for *Inquiry* and non-*Inquiry* students, as the difference in model deviance between a model with a step structure that varies based on curriculum and a model with an invariant step structure across curricula is not statistically significant (χ^2 difference = 48.62, df = 44, p = 0.29).

Relationship between Inquiry Project curriculum participation, grade level and student performance. Within-classroom variation was examined across four subgroups: Inquiry Project participants, Inquiry Project non-participants, Grade 4 students, and Grade 5 students.

Scale, Proportion, and Quantity dimension.

Within-classroom variation among Inquiry Project and non-Inquiry students. Figure 4.27 contains boxplots of student ability estimates, grouped by classroom

membership. There were 18 classrooms in the pilot sample. The composition of the Inquiry and non-Inquiry samples are dissimilar, because there are many more 4th grade students in the Inquiry sample than the non-Inquiry sample. When the analysis is limited to 5th graders, 11 classrooms remain. There are still some visible differences in variability between classrooms, but overall differences are reduced (Figure 4.28).

Furthermore, if classroom variance estimates are compared (Figure 4.29), it appears that fifth-grade *Inquiry Project* classrooms tended to have very similar variance in comparison to more diverse classroom variance estimates among the non-*Inquiry* classrooms. Note the unequal number of classrooms between groups, which may limit comparisons.

Average ability on the Scale, Proportion, and Quantity dimension was higher in *Inquiry Project* classrooms (fifth-grade *Inquiry Project* mean = 0.36 logits, fifth-grade non-*Inquiry* mean = -0.26 logits, $t = 3.71$, $df = 206$)[33].

Within-classroom variation among fourth- and fifth-grade students. Refer to Figure 4.30 for a visual representation of within-classroom variability among all classrooms. The vast majority of 4th grade students were *Inquiry Project* participants, which made the composition of the 4th and 5th grade samples unequal. Therefore, the analysis was limited to *Inquiry Project* classrooms to compare variability within classrooms of different grade levels (Figure 4.30).

Differences in within-classroom variability still remain after the analysis is constrained to *Inquiry Project* participants, but these differences do not appear to be

[33] Note that this result does not account for similarities among students in the same classroom, therefore, the standard error may be underestimated.

associated with grade level. For example, classrooms 6 and 8, which have the largest and smallest interquartile ranges, respectively, are both 4th grade classrooms.

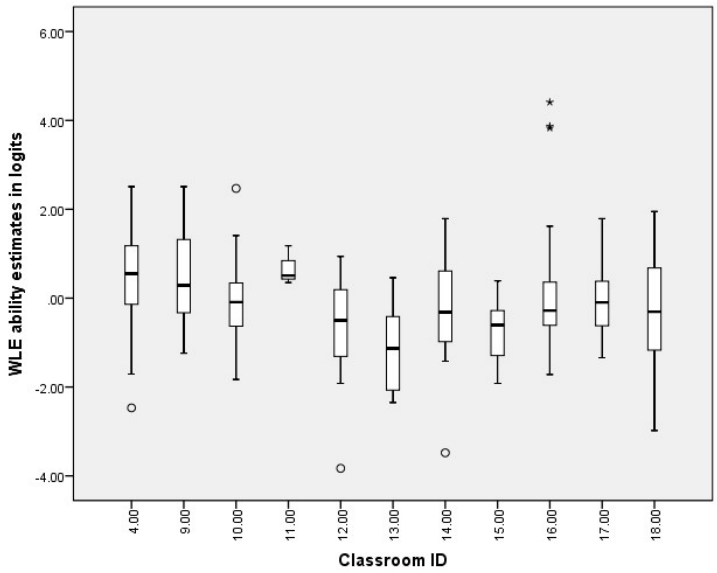


Figure 4.28. Student ability boxplots, grouped by fifth-grade classroom, for the Scale, Proportion, and Quantity dimension. Classrooms 4, 9, and 10 were Inquiry Project participants, while classrooms 11-18 were not.

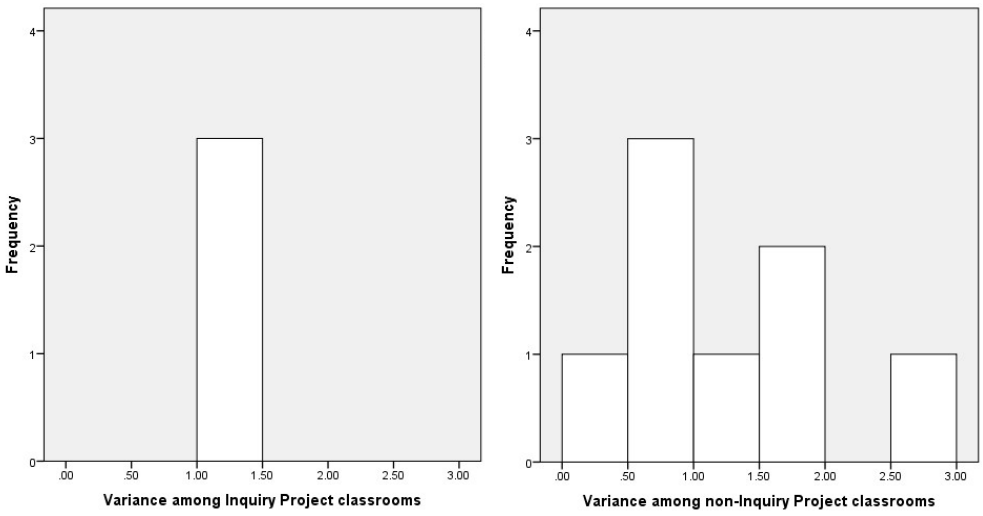


Figure 4.29. Histograms of within-classroom variance in student ability for the Scale, Proportion, and Quantity dimension. Variance estimates for *Inquiry Project* classrooms are clustered between 1.0 and 1.50, while variance estimates among non-*Inquiry* classrooms range from 0.0 to 3.0.

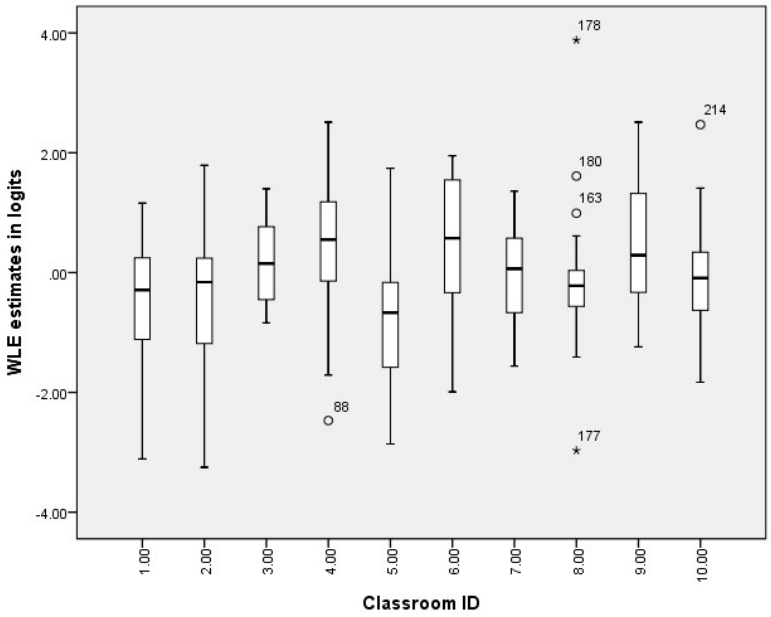


Figure 4.30. Student ability boxplots, grouped by *Inquiry Project* classroom, for the Scale, Proportion, and Quantity dimension. Classrooms 4, 9, and 10 were fifth-grade classes, while the remaining class rooms were fourth-graders.

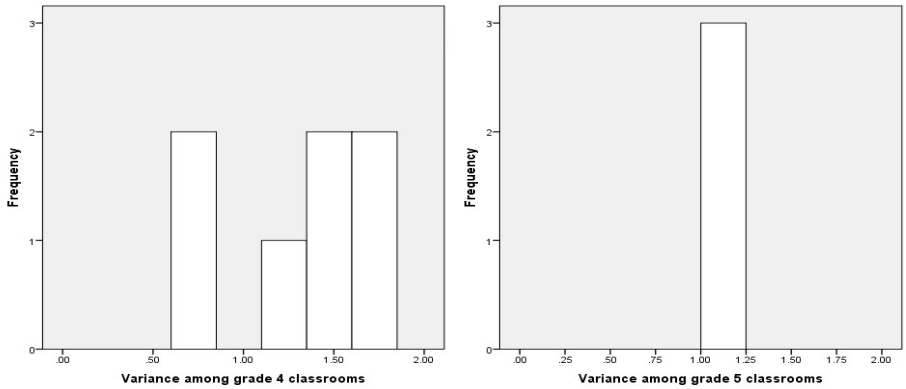


Figure 4.31. Histograms of within-classroom variance in student ability for the Scale, Proportion, and Quantity dimension. Variance estimates for 5th grade classrooms are clustered between 1.00 and 1.25, while variance estimates among 4th grade classrooms range from 0.50 to 1.75.

If classroom variance estimates are compared (Figure 4.31), it appears that fifth-grade classrooms tended to have very similar variance in comparison to more diverse

classroom variance estimates among the fourth-grade classrooms. Note the unequal number of classrooms between groups, which may limit comparisons.

Average ability on the Scale, Proportion, and Quantity dimension was higher among grade 5 students (fifth-grade *Inquiry Project* mean = 0.36 logits, fourth-grade *Inquiry Project* mean = -0.19 logits, t = -3.47, df = 216)[34].

Structure and Properties of Matter.

Within-classroom variation among Inquiry Project and non-Inquiry students. Visually, a similar pattern emerges for the Structure and Properties of Matter dimension. When examining all classrooms, there does appear to be some variation in the amount of within-group variability. Classrooms 6 and 15 have the smallest interquartile range of the 18 groups, about half of a logit, where classrooms 12 and 13, for example, have much larger spreads (Figure 4.32). When the analysis is limited to 5th graders, 11 classrooms remain. There are still some visible differences in variability between classrooms, note especially classrooms 11 and 12, which have very different ranges in student performance (Figure 4.33).

Again, histograms (Figure 4.34) revealed that *Inquiry Project* classrooms tended to have more similar variance, while non-*Inquiry* classrooms had much more diversity in variance. Overall, there are no consistent trends in the amount of variance in *Inquiry Project* classrooms and non-*Inquiry* classrooms.

[34] Note that this result does not account for similarities among students in the same classroom, therefore, the standard error may be underestimated.

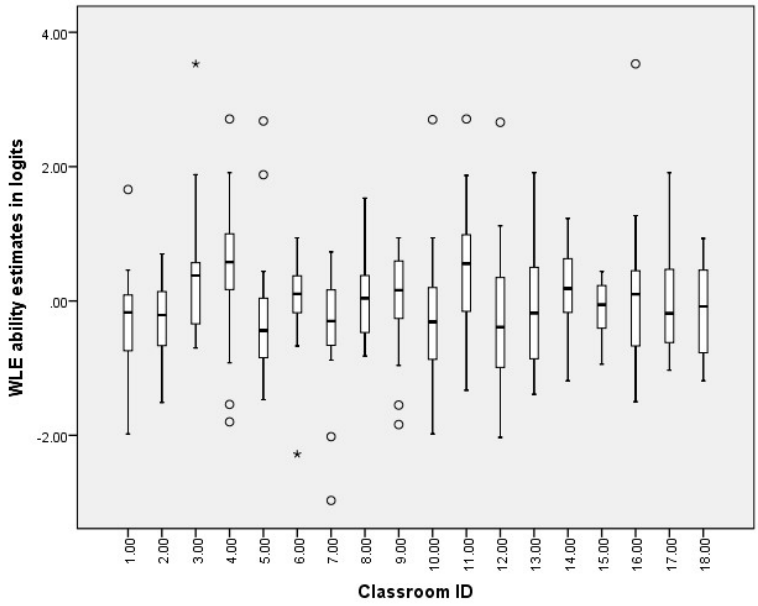


Figure 4.32. Student ability boxplots, grouped by classroom, for the Structure and Properties of Matter dimension. Classrooms 1-10 were *Inquiry Project* participants, while classrooms 11-18 were not. Classrooms 4, 9, 10, 12, and 14-18 are 5th grade classrooms, classrooms 1-3 and 5-8 are 4th grade classrooms, and classrooms 11 and 13 are mixed classrooms.

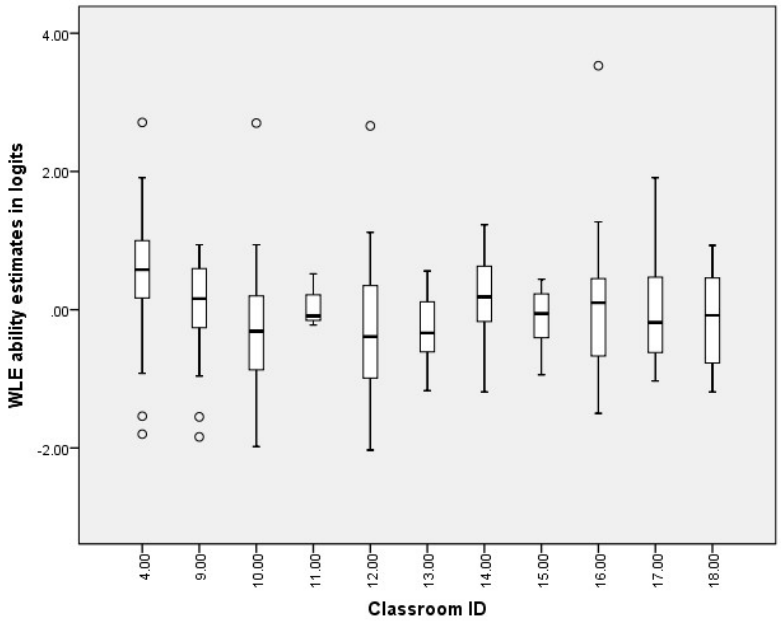


Figure 4.33. Student ability boxplots, grouped by fifth-grade classroom, for the Structure and Properties of Matter dimension. Classrooms 4, 9, and 10 were *Inquiry Project* participants, while classrooms 11-18 were not.

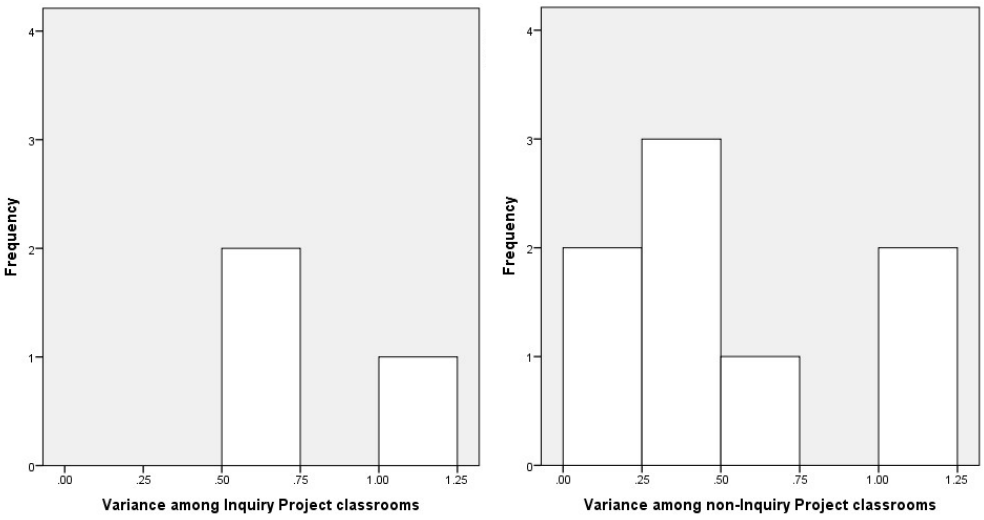


Figure 4.34. Histograms of within-classroom variance in student ability for the Structure and Properties of Matter dimension. Variance estimates for non-*Inquiry* classrooms range from 0.00 to 1.24, while *Inquiry Project* classrooms fall between 0.50 and 1.25.

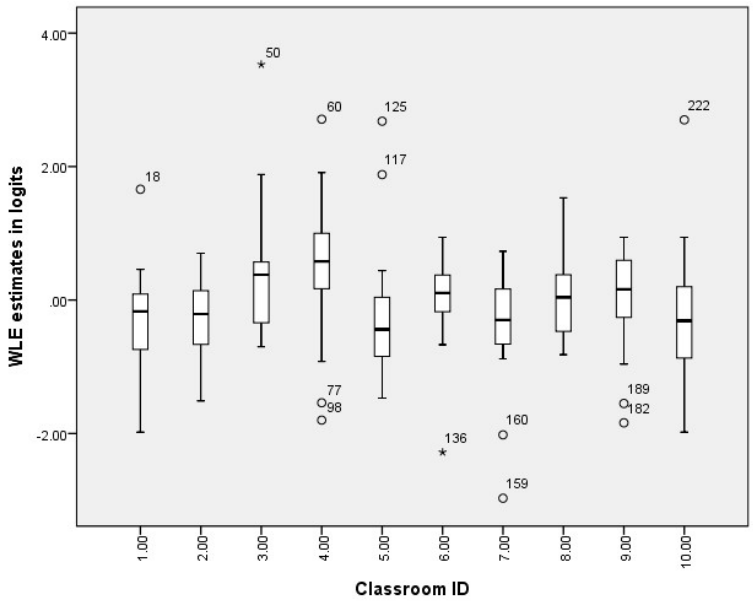


Figure 4.35. Student ability boxplots, grouped by *Inquiry Project* classroom, for the Structure and Properties of Matter dimension. Classrooms 4, 9, and 10 were fifth-grade classes, while the remaining class rooms were fourth-graders.

Average ability on the Structure and Properties of Matter dimension was higher in *Inquiry Project* classrooms (fifth-grade *Inquiry Project* mean = 0.24 logits, fifth-grade non-*Inquiry* mean = -0.06 logits, t = 2.55, df = 206)[35].

Within-classroom variation among fourth- and fifth-grade students. Refer to Figure 4.32 for a visual representation of within-classroom variability among all classrooms. To compare differences between fourth- and fifth-grade classrooms, the analysis was again limited to *Inquiry Project* classrooms (Figure 4.35). Differences in within-classroom variability largely disappear after the analysis is constrained to *Inquiry Project* participants, with many classrooms appearing to have very similar distributions of student ability.

If classroom variance estimates are compared (Figure 4.36), it appears that fifth-grade classrooms tended to have more similar variances in comparison to more diverse classroom variance estimates of the fourth-grade classrooms. Note the unequal number of classrooms between groups, which may limit comparisons.

Average ability on the Scale, Proportion, and Quantity dimension was higher among grade 5 students (fifth-grade *Inquiry Project* mean = 0.24 logits, fourth-grade *Inquiry Project* mean = -0.13 logits, t = -3.06, df = 216)[36].

Engaging in Argument from Evidence.

Within-classroom variation among Inquiry Project and non-Inquiry students. On the Engaging in Argument from Evidence dimension, variance within classrooms appears to be somewhat more uniform than the variances in the other two dimensions. There are

[35] Note that this result does not account for similarities among students in the same classroom, therefore, the standard error may be underestimated.
[36] Note that this result does not account for similarities among students in the same classroom, therefore, the standard error may be underestimated.

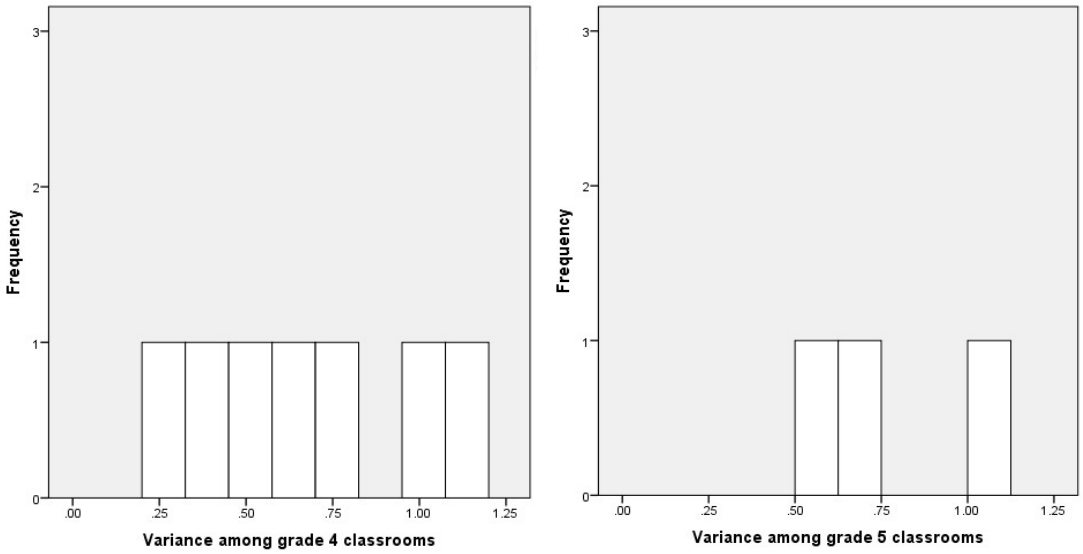


Figure 4.36. Histograms of within-classroom variance in student ability for the Structure and Properties of Matter dimension. Variance estimates for 5th grade classrooms are clustered between 0.50 and 1.13, while variance estimates among 4th grade classrooms range from 0.25 to 1.25.

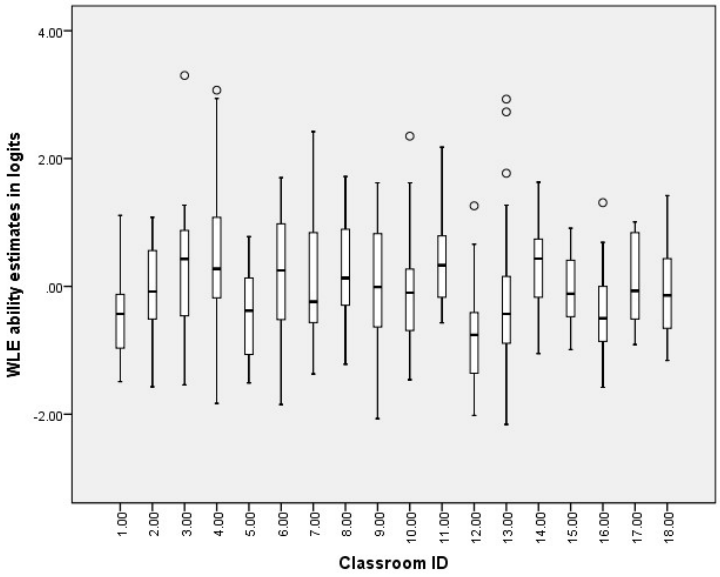


Figure 4.37. Student ability boxplots, grouped by classroom, for the Engaging in Argument from Evidence dimension. Classrooms 1-10 were Inquiry Project participants, while classrooms 11-18 were not. Classrooms 4, 9, 10, 12, and 14-18 are 5th grade classrooms, classrooms 1-3 and 5-8 are 4th grade classrooms, and classrooms 11 and 13 are mixed classrooms.

discrepancies between some classrooms (e.g., the ranges of classrooms 4 and 5 are quite different), but for the most part, the interquartile ranges are similar across classrooms (Figure 4.37). When the analysis is limited to 5th graders, differences in within-classroom variability become more pronounced (Figure 4.38). In particular, the 3 *Inquiry Project* classrooms (classrooms 4, 9, and 10) tend to demonstrate larger variance in student performance than the remaining non-*Inquiry* classrooms.

Again, histograms (Figure 4.39) revealed that *Inquiry Project* classrooms tended to have more similar variances in comparison to those non-*Inquiry* classrooms. Average ability on the Engaging in Argument from Evidence dimension was higher in *Inquiry*

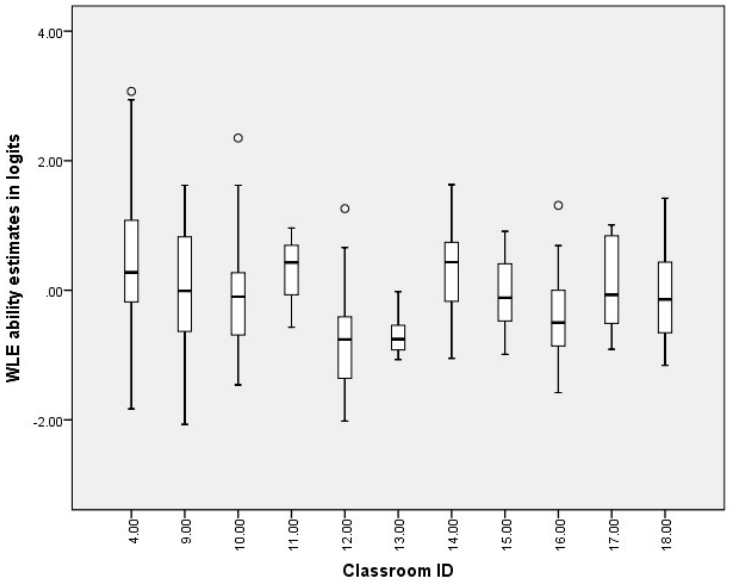


Figure 4.38. Student ability boxplots, grouped by fifth-grade classroom, for the Engaging in Argument from Evidence dimension. Classrooms 4, 9, and 10 were *Inquiry Project* participants, while classrooms 11-18 were not.

Project classrooms (fifth-grade *Inquiry Project* mean = 0.30 logits, fifth-grade non-*Inquiry* mean = -0.20 logits, t = 3.89, df = 206)[37].

Within-classroom variation among fourth- and fifth-grade students. Refer to Figure 4.37 for a visual representation of within-classroom variability among all classrooms. Again, the analysis was limited to *Inquiry Project* classrooms to examine differences between grade levels (Figure 4.40). Differences in within-classroom variability still remain after the analysis is constrained to *Inquiry Project* participants, but these differences do not appear to be associated with grade level. The interquartile range of grade 4 classrooms appears to vary as much as the interquartile range of grade 5 classrooms.

If classroom variance estimates are compared (Figure 4.41), it appears that fifth-grade classrooms tended to have more similar variances relative to the more diverse classroom variance estimates of fourth-grade classrooms. Note the unequal number of classrooms between groups, which may limit comparisons.

Average ability on the Scale, Proportion, and Quantity dimension was higher among grade 5 students (fifth-grade *Inquiry Project* mean = 0.30 logits, fourth-grade *Inquiry Project* mean = -0.04 logits, t = -2.52, df = 216)[38].

[37] Note that this result does not account for similarities among students in the same classroom, therefore, the standard error may be underestimated.
[38] Note that this result does not account for similarities among students in the same classroom, therefore, the standard error may be underestimated.

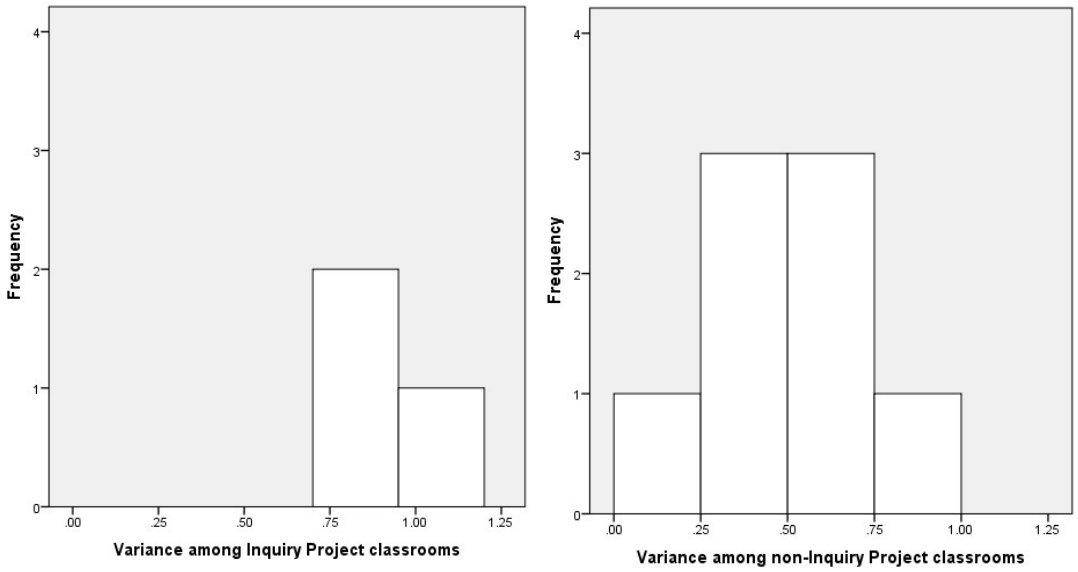


Figure 4.39. Histograms of within-classroom variance in student ability for the Engaging in Argument from Evidence dimension. Variance estimates for Inquiry Project classrooms are clustered between 0.75 and 1.25, while variance estimates among non-Inquiry classrooms range from 0 to 1.

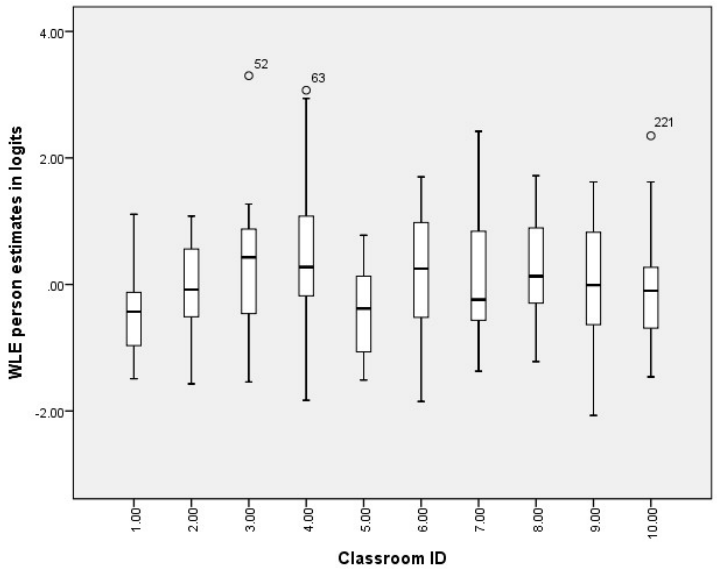


Figure 4.40. Student ability boxplots, grouped by *Inquiry Project* classroom, for the Engaging in Argument from Evidence dimension. Classrooms 4, 9, and 10 were fifth-grade classes, while the remaining class rooms were fourth-graders.

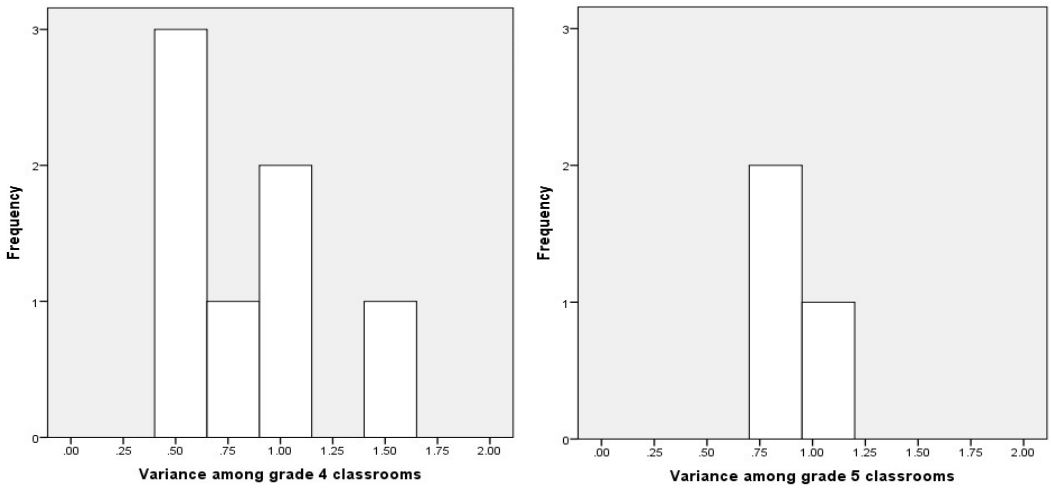


Figure 4.41. Histograms of within-classroom variance in student ability for the Structure and Properties of Matter dimension. Variance estimates for 5th grade classrooms are clustered between 0.75 and 1.13, while variance estimates among 4th grade classrooms range from 0.40 to 1.60.

HLM analysis. The previous analyses examined whether two group characteristics – participation in the Inquiry Project curriculum, and grade level – were associated with within-classroom variability in student performance on each of the three assessment dimensions. However, these analyses did not account for similarities among students who share a teacher. Because individual students within a classroom share a common context for learning, the amount of within-classroom variance may be underestimated. Therefore, a hierarchical linear model was employed to reexamine the amount of variance shared among students in a classroom.

Three unconditional level-1 models were generated with student ability estimates on the three assessment dimensions as the outcome variables. The unconditional model contains no individual or group predictors of student achievement, but simply allows the intercept to vary randomly based on classroom membership. This allows for partitioning of variance into within-group and between-group sources and the calculation of the ICC,

or the percentage of variance in the outcome that is attributable to classroom membership.

Within-group and between-group variance estimates, calculated by HLM, are presented for each of the following subgroups: 4th grade *Inquiry Project* students, 5th grade *Inquiry Project* students, and 5th grade non-*Inquiry Project* students (Table 4.25).

To directly compare the within-group variance of independent groups, an F-test was employed, with the degrees of freedom adjusted to account for the loss of information caused by students sharing a common context. Results for the F-test are presented in Table 4.26.

Table 4.25.

Variance Estimates from Unconditional Multilevel Models

	Variance between groups	Variance within groups	Total variance	Intraclass correlation coefficient (ICC)
Scale, Proportion, and Quantity (WLE estimate)				
Inquiry Project G5	0.07	1.15	1.22	0.06
Inquiry Project G4	0.12	1.33	1.45	0.08
non-*Inquiry Project* G5	0.08	1.58	1.65	0.05
Structure and Properties of Matter (WLE estimate)				
Inquiry Project G5	0.18	0.75	0.93	0.19
Inquiry Project G4	0.03	0.68	0.71	0.05
non-*Inquiry Project* G5	0.00	0.65	0.65	0.00
Engaging in Argument from Evidence (WLE estimate)				
Inquiry Project G5	0.06	1.05	1.11	0.05
Inquiry Project G4	0.05	0.80	0.85	0.06
non-*Inquiry Project* G5	0.12	0.54	0.66	0.18

Table 4.26.

Results of F-test Comparing Within-Classroom Variances Between Subgroups

Scale, Proportion, and Quantity (WLE estimate)					
Within-group variance	df	Within-group variance	df		
Inquiry Project G5		*Inquiry Project* G4		F	p-value
1.15	86	1.33	120	0.87	0.48
Inquiry Project G4		non-*Inquiry Project* G5			
1.33	120	1.58	109	0.84	0.36
Inquiry Project G5		non-*Inquiry Project* G5			
1.15	86	1.58	109	0.73	0.13
Structure and Properties of Matter (WLE estimate)					
Within-group variance	df	Within-group variance	df		
Inquiry Project G5		*Inquiry Project* G4		F	p-value
0.75	86	0.68	120	1.10	0.62
Inquiry Project G4		non-*Inquiry Project* G5			
0.68	120	0.65	109	1.05	0.80
Inquiry Project G5		non-*Inquiry Project* G5			
0.75	86	0.65	109	1.16	0.47
Engaging in Argument from Evidence (WLE estimate)					
Within-group variance	df	Within-group variance	df		
Inquiry Project G5		*Inquiry Project* G4		F	p-value
1.05	86	0.80	120	1.31	0.17
Inquiry Project G4		non-*Inquiry Project* G5			
0.80	120	0.54	109	1.48	0.04
Inquiry Project G5		non-*Inquiry Project* G5			
1.05	86	0.54	109	1.94	0.00

On the Scale, Proportion, and Quantity dimension, within-classroom variance is lowest among Grade 5 students who took the *Inquiry Project* curriculum, and highest among Grade 5 students who did not take the *Inquiry Project* curriculum. However, none of these differences are statistically significant, suggesting that curriculum exposure and grade level do not affect the amount of variability in student performance within a classroom. This aligns with the previous finding based on the uncorrected variance

estimates. Additionally, the ICC for all groups is below 0.10, suggesting that there is not much difference in performance attributable to classroom membership.

On the Structure and Properties of Matter dimension, within-classroom variance is lowest among Grade 5 students who did not take the *Inquiry Project* curriculum, and highest among Grade 5 students who took the *Inquiry Project* curriculum. However, none of these differences are statistically significant, suggesting that classroom exposure and grade level do not affect the amount of variability in student performance within a classroom. This aligns with the previous finding based on the uncorrected variance estimates. Among Grade 5 students in the *Inquiry Project* curriculum, the ICC is much higher than the other two groups; 19% of variation is student performance is attributable to classroom membership. Differences between classrooms are not present at all in the non-*Inquiry Project* sample, and to a much lesser extent among Grade 4 *Inquiry Project* students.

On the Engaging in Argument from Evidence dimension, within classroom variance is highest among Grade 5 *Inquiry Project* students, and lowest among Grade 5 non-*Inquiry Project* students. Differences between these two groups are statistically significant, suggesting that there is significantly more variability in Grade 5 classrooms where student took the *Inquiry Project* curriculum, compared to Grade 5 classrooms where students did not participate in the curriculum. The difference between Grade 4 *Inquiry Project* students and Grade 5 non-*Inquiry Project* students is also statistically significant. Again, *Inquiry Project* classrooms demonstrated more within-class variability than those that did not participate in the curriculum. The difference between the Grade 4 and Grade 5 *Inquiry Project* groups was not statistically significant. These findings align

with the previously reporting findings based on uncorrected variance estimates, and suggest that classrooms who participate in the *Inquiry Project* curriculum have a larger range of student performance within classrooms on the Engaging in Argument from Evidence dimension than classrooms who do not participate in the curriculum. Student performance in both *Inquiry Project* groups is higher than the non-*Inquiry Project* group, on average (G5 *Inquiry Project* mean = 0.30, G4 *Inquiry Project* mean = -0.04, G5 non-*Inquiry Project* mean = -0.20). Finally, the ICC is much higher among non-*Inquiry Project* classrooms than among *Inquiry Project* classrooms; 18% of variability in student performance is attributable to classroom membership. Differences between classrooms account for a much smaller percentage of variability in student performance among students who participated in the *Inquiry Project* curriculum.

Chapter 5: Conclusion

With the introduction of the *Next Generation Science Standards* (NGSS Lead States, 2013) as the preeminent framework for K-12 science education, there is a need for assessment to support and guide implementation of the new multidimensional standards in classrooms. Multidimensional assessment is a particular challenge for assessment developers, especially in the area of science, where assessment has usually focused on content to the exclusion of practice (Pellegrino, 2013). There are several aspects to this challenge, including multidimensional task design (Gorin & Mislevy, 2013), and the identification of appropriate psychometric methods to locate student ability on multiple related constructs (Pellegrino, et al., 2014).

The purpose of the study was to explore item design and psychometric considerations for multidimensional science assessment in the context of an elementary assessment of the 4th grade *Inquiry Project* curriculum (TERC, 2011). The study examined 3 factors that might affect the validity and interpretation of the resulting estimates of multidimensional student performance: item design (scaffolding and response format), scoring, and psychometric modeling of student performance data. The study also examined some general indicators of assessment quality and validity, and provided some small insight into the effectiveness of the *Inquiry Project* curriculum. The following chapter includes a summary of the primary research findings, implications for multidimensional assessment, study limitations, and directions for future research.

Summary of findings

Research question 1: Multidimensional scaffolding – what is its effect on student responses?

Scaffolding helps students attend to critical pieces of assessment tasks. Students tended to overlook critical pieces of information more frequently when less scaffolding was used.[39] Specifically, they were more likely to overlook information when a single-prompt explicit multidimensional item format was used. For a single-prompt implicit multidimensional task, the stimulus information and task prompt draws on many dimensions of science learning, but only presents the student with one question/task. It seems likely that the density of questions/information in single-prompt multidimensional items was overwhelming for novice science learners, making it more likely that they would overlook something important. On the other hand, multiple-prompt explicit multidimensional present information to students in a more controlled way. This hypothesis is supported by observations from the students. For example, one student was so overwhelmed by the lengthy prompt in a single-prompt unidimensional item that he scaffolded the task himself by numbering each part of the task and addressing each piece separately, in order to avoid accidentally overlooking part of the task.

There does not appear to be a relationship between the amount of multidimensional scaffolding and student understanding of the nature of the subtasks. There were no clear changes in the frequency with which students expressed difficulty

[39] It must be noted that overlooking a critical piece of information was observed only 7 times during the first round of cognitive interviews, making it difficult to draw definitive conclusions about the effect of scaffolding on the likelihood of observing this issue. However, this is a plausible mechanism for affecting students' experience with the items, and one which should be monitored as other characteristics of student performance are examined.

understanding a subtask or confusion about the item layout, depending on how much scaffolding was used. Overall, this suggests that the amount of scaffolding may not affect student understanding of the nature of the task, or at least that it may not affect their understanding in a way that the student is able to perceive and express verbally.

Scaffolding is associated with more thorough, informative responses. Item variations with more scaffolding produce student responses that are more informative about their understanding of science. Data from both the cognitive labs and the pilot tests demonstrated that students' responses addressed more of the assessment dimensions, on average, when they responded to items with more scaffolding. As the amount of scaffolding decreased, the students' answers tended to provide less information about the students' science ability across all three dimensions. Furthermore, students were more likely to provide information relevant to all three dimensions when presented with a multiple-prompt item. This difference was most notable on the Scale, Proportion, and Quantity, and Structure and Properties of Matter; when a single-prompt item was used, students were more likely to give missing responses. Students were more likely to provide explicit documentation of their reasoning about measurement and matter-related concepts when they were explicitly and separately prompted to do so.

The Engaging in Argument from Evidence dimension was less affected by the amount of item scaffolding. This finding may reflect the nature of a single-prompt item, which naturally lends itself to an extended response with reasoning. The multiple-prompt format does not provide as much extra emphasis and structure for argumentation, as it does for the other content-based dimensions. This may explain why no difference is observed due to scaffolding.

One of the most striking findings was that students were more likely to leave single-prompt items entirely blank than multiple-prompt items. The addition of scaffolding appears to make items more accessible to students, above and beyond directing them to attend to all dimensions. On the surface, students may find open-ended items with large response spaces to be intimidating. Thus, students are more likely to bypass single-prompt items altogether, compared to multiple-prompt items where the response space is broken up into several pieces. The use of multiple-prompt items appears to increase the total amount of information gathered from each student in two ways: by reducing the entry barrier for providing *any* response, and by reinforcing student attention and response to each dimension separately.

Credible inferences based on assessment results are dependent on the amount of relevant evidence generated – a claim is made about a student based on their responses (Messick, 1988; Mislevy, 2007) – and without sufficient relevant evidence, the inference lacks validity. Finding ways to increase the amount and quality of evidence is thus a high priority for assessment developers. It appears that scaffolding is an effective way to increase the amount of evidence provided by the student, especially when faced with complex multidimensional tasks.

Scaffolding is most helpful for lower-ability students. The association between scaffolding and missing responses varied depending on student ability. The difference in the extent of missing responses between single-prompt and multiple-prompt items was much larger among students with low estimated ability than among students with high estimated ability. This was particularly true on the Scale, Proportion, and Quantity and Structure and Properties of Matter dimensions. It is likely that the additional scaffolding

provided in the multiple-prompt item format affects response rates by making the item accessible to a larger number of students. For lower ability students, the single-prompt items were more difficult to access, and student responses were less likely to address all of the assessment dimensions. These students were much more likely to provide a relevant response with multidimensional scaffolding. Higher ability students found the single-prompt items more accessible, and the addition of multidimensional scaffolding did not provide as much of a "bump" as it did for the lower-ability students. This finding is in line with other research demonstrating that assessment scaffolding is more beneficial to students with lower ability (Gotwals & Songer, 2013).

There were sizeable differences in the number of blank responses between single-prompt and multiple-prompt items among all ability groups (low, medium, and high) when ability is grouped based on the Scale, Proportion, and Quantity and Structure and Properties of Matter dimensions. This suggests that whatever feature of the single-prompt format that inhibits its accessibility affects all students, regardless of their underlying ability.

The relationship between student ability and the effect of scaffolding was less clear on the Engaging in Argument from Evidence dimension. Again, the nature of a single-prompt item lends itself to argumentation, which may lessen the impact of extra scaffolding, even among low ability students.

Scaffolding requires more response time. Items with more multidimensional scaffolding tend to require more response time from students, on average. Since items with more multidimensional scaffolding tended to collect more information relevant to all three assessment dimensions, this finding is unsurprising. Classroom time is valuable,

however, and the benefit of collecting more thorough information about student science understanding should be weighed against the drawback of increasing the amount of time students spend on the assessment.

Alignment between the item structure and scoring rubric is essential. Greater amounts of multidimensional scaffolding were associated with higher average interrater reliability when a multidimensional scoring rubric was used, but the opposite pattern was observed when a holistic rubric was used. When the interrater reliability of item variations with different amounts of scaffolding were directly compared, the results were less clear, but tended to follow the same pattern. When items were designed to elicit separate, explicit evidence of student understanding for each dimension, it was easier for raters to agree upon scores when the rubric assigned separate multidimensional scores to each piece of student evidence. On the other hand, when the item structure required students to integrate their understanding into a single extended response, it was easier for raters to agree upon scores when the rubric assigned a single holistic score to the response. This suggests that deliberate alignment between the item structure and the structure of the scoring rubric facilitates greater ease of scoring, which is reflected in the extent of agreement between raters. This further reinforces the notion that all parts of the assessment process (construct, items, rubrics, and measurement models) should be engineered together, aligned with each other and with the common purpose of the assessment (Brown & Wilson, 2011), even to the extent that an item's response structure is reflected in its scoring rubric.

The relationship between item difficulty and amount of scaffolding varies by dimension. There was no clear association between scaffolding and difficulty among

Scale, Proportion, and Quantity items. This finding held for the overall group, and for a subgroup analysis in which item difficulties were calculated separately for low, medium, and high ability students. Sometimes single-prompt items were easier, and sometimes multiple-prompt items were easier. If there is a scaffolding effect, it does not seem to be consistent. It is possible that scaffolding's effect varies depending on the specific item content. For instance, one of the item variations (ROMITA) asks students to make a judgment about the weight of two objects, and then use that judgment to speculate about whether or not they might be made of the same material. In the multiple-prompt version of the item, students are asked to state precisely which item (if any) is heavier. In the single-prompt version they are only asked to make a judgment about the objects' material, presumably by drawing on information about the items' weight. Because the multiple-prompt item specifically asks for students to make an assertion about the weight of the objects, it may call out a misconception that weight can be measured by making judgments based on what can be felt with the senses. This misconception can be masked by other aspects of the student's response in the single-prompt version of the item, thereby making it more likely that a rater will misjudge the student's understanding of weight measurement when scoring the single-prompt variation. The Scale, Proportion, and Quantity component of the multiple-prompt version of ROMITA was more difficult than the single-prompt version, lending credence to this explanation. However, this pattern was not observed consistently – possibly because the remaining items included more straightforward measurement opportunities, or because item difficulty was related to other confounding factors or due to chance.

On the Structure and Properties of Matter dimension, single-prompt items tend to be easier than multiple-prompt items, and this pattern held for students of all abilities. There are two different explanations for this difference, and it is likely that both explanations play a role. One is that – like the Scale, Proportion, and Quantity explanation presented above – splitting the item into multiple, targeted prompts forces students to grapple with certain aspects of the task that they may have otherwise avoided or brushed over. For example, SUGAR, when split up into multiple prompts, specifically asked students to consider whether a grain of sugar weighs anything at all, in addition to asking them to calculate the weight. The single-prompt version prompted the weight calculation only, which highlighted the Scale, Proportion, and Quantity dimension. It was assumed that any student misconceptions about whether or not a grain of sugar has *any* weight would manifest in their answer as a response of "zero" or some other indication that a single grain has no weight. Likewise, it was assumed that students who correctly performed the measurement calculation also understood the matter concept. This assumption may not be tenable, especially since the single-prompt formulation emphasizes the measurement task; this version of the item may enable students to neglect the overall consideration of whether weight is even present. Similarly, another item (ANA) asks students to consider the weight of clay before and after it has been shaped into a ball. In the multiple-prompt version, students are specifically prompted to consider whether the weight is the same or different, while the single-prompt version simply asks them to provide the weight of the clay ball and support with evidence – again, highlighting the measurement aspect of the question and leaving the matter aspect implicit. The multiple-prompt version of BOX forces students to consider pairs of items

based on size, weight, and density before asking which might be made of the same material, whereas the single-prompt item only requires them to make a judgment about which blocks might be made of the same material. When combined into a single prompt, students may focus on only one main aspect of the item, since there is only one response space. In these items, the measurement task is the most salient aspect of the single-prompt item. Splitting the item forces students to address the dimensions separately, which forces them to confront and address misconceptions that they may have brushed over in their response to a single prompt.

Alternatively, the single-prompt format may make raters more susceptible to a halo effect, in which they brush over ambiguities in student responses by assuming that the student's understanding on one dimension is in line with their performance on the other two dimensions. It seems reasonable to expect that raters may have difficulty taking notice of students' weaknesses when their responses confound several related skills, any of which may be stronger or weaker than the others for a particular student. For example, a student's response to the single-prompt version of SUGAR may reflect a student's proficiency with the concept of proportionality. Because the student demonstrates proficiency with one of the skills assessed, it may lead the rater to presume that the student also has proficiency with other skills – in this case, they may assume that the student recognizes the weight of a grain of sugar. However, this is not necessarily a valid assumption. If this assumption is left unchecked across many students, it may artificially shift the difficulty estimates for the single-prompt versions of the items.

On the Engaging in Argument from Evidence dimension, there is no clear relationship between item difficulty and response format. The overall item estimates

based on the full dataset suggest that multiple-prompt item thresholds may have a larger range of difficulty than the single-prompt thresholds. However, this pattern is not upheld among smaller subgroups of low, medium, and high ability students. The subgroup analysis suggests that the lower thresholds tend to be more difficult for low ability students when the single-prompt format is used, compared to the multiple-prompt format. Among medium and high ability students, this gap disappears. Providing even a weak argument is more difficult for low ability students when a single-prompt format is used, suggesting that low-ability students struggle more with issues of access when compared to their higher ability counterparts. Otherwise, the pattern of item difficulty on the Engaging in Argument from Evidence dimension is muddled and difficult to interpret, indicating that scaffolding is probably not responsible for any observed differences.

Multidimensional scaffolding is not associated with item fit. Item fit statistics did not reveal any patterns related to the amount of scaffolding used in an item. This may have been due to the small number of items, or the fact that most items fit the model well.

Overall conclusions about scaffolding. Scaffolding offers several clear advantages for the assessment of multidimensional constructs. Scaffolding helps direct student attention to critical features of the assessment task. This results in student responses that contain better evidence of student ability than items without scaffolding. Scaffolding enables lower-ability students to access assessment tasks that they may have otherwise ignored, thus providing higher quality diagnostic information about these students' particular weaknesses, which is useful in supporting instruction. Scaffolding may also benefit raters, who are better able to make judgments about student ability when scaffolding helps structure responses so that they are explicitly aligned with scoring

criteria. Overall, scaffolding seems to greatly enhance assessment validity by providing raters with more, higher quality evidence of student performance, in a format that is easier for raters to digest and evaluate.

There are some drawbacks. Highly scaffolded items require more testing time from students, a factor that must be weighed against the additional information that they provide. But more importantly, the NGSS present the three dimensions of science learning as components of an *integrated* understanding of science. Scaffolding, it can be argued, forces a degree of separation between these dimensions; by splitting the item into separate prompts, students are able to focus on each dimension in isolation. Single-prompt items are more authentic, in that they require students to actively integrate their relevant knowledge/skills from these dimensions on their own. Furthermore, when students address the three dimensions in a scaffolded item, we cannot be sure this is because the student naturally conceives of multiple dimensions that reinforce each other. The single-prompt item may be informative in its own right; the missing aspects of a response are evidence of students' failure to integrate, which is a weakness in their understanding of the nature of science. Scaffolded items may provide better evidence of student performance on the three dimensions individually, but when it comes to representation of science as an *integrated* multidimensional construct, some validity concerns remain.

Research question 2: Response format – what is its effect on student responses?

The selected-response format was more likely to introduce construct-irrelevant variance and confusion. There appears to be a difference in student understanding of the

intended task, depending on whether they interact with a constructed-response or a selected-response version of an item. Students were far more likely to express confusion about how to complete a task when they were given a selected-response version of the task. This is largely due the argument subtask. For a selected-response argument, students were given a list of potential pieces of evidence and reasoning, and asked to select those that best explained their answer. Students frequently asked for clarification about how to perform the selected-response version of this task. Many said that they had never seen a question like that before. Some students asked for clarification about how to respond to the question, and how many answer options they were allowed to choose. A few students missed the point of the question, and simply selected all responses that were true, based on the scenario, rather than only the subset of answer options that supported their answer. Therefore, it seems that without previous exposure or test preparation, the selected-response items, and the selected-response arguments in particular, are prone to student misinterpretation.

Similarly, there were some differences between the intended and enacted construct, depending on response format. When faced with a selected-response item, students were likely to complain that the answer options didn't reflect their understanding. These students ended up choosing the answer options that were closest to their understanding, but were not perfect representations of their mental model. This creates some distance between the item's intended construct and the enacted construct. To avoid this problem, more response options can be offered to reflect a greater variety in student understanding. However, there were several items with a large number of possible responses, with the intent of meeting this purpose. Selected-response versions of

the argument subtask, in particular, contained lengthy reading intensive response options, and often 10 or more – more than twice the standard number. Students expressed dislike for these items, as they were overwhelmed by the large amount of text/information. Regarding the appropriate number of response options, the set of constraints makes it difficult to find an optimal solution, begging the question of whether the selected-response format is appropriate for such complex multidimensional items.

Many students express a preference for the selected-response format. It must be noted that the majority of students stated a preference for the selected-response versions of the items over the constructed-response versions, despite any complaints or misunderstandings they may have exhibited with the format. When asked why they preferred the selected-response versions of the items, students usually stated that the answer options provided a good starting point for thinking about the items, and helped them to articulate their reasoning in ways that they might not have been able to express otherwise. Several students reported that the answer options led them to reconsider their answer, usually by suggesting compelling alternative reasoning. This suggests that the answer options may provide a form of additional scaffolding that is helpful for students who are on the cusp of understanding, consistent with an understanding of scaffolding as an aid for students who are within the *zone of proximal development* (Vygotsky, 1978).

Response options influence the content of student answers. The selected-response format affected the response given by the students. When students were asked to write an argument and then select pieces of evidence and reasoning in a selected-response argument, they supplemented their written responses with additional evidence and reasoning about half of the time. A few students completely changed their answers after

seeing the response options (11% and 14% of the time, depending on the specific item variation). Finally, some students chose not to provide a written response, but were able to choose from among the response options. Based on these observations and the number of students who verbally admitted to rethinking responses after seeing response options, it appears that response options may influence the way students think about a question, sometimes leading them to different or more detailed conclusions than they would have reached otherwise.

When given a selected-response version of an argument, students tended to choose a greater amount of evidence and reasoning than was reflected in their written arguments. In particular, they tended to choose a greater amount of both relevant-supporting evidence and irrelevant/unsupportive evidence than they provided in their written responses. This can be triangulated by observations from the cognitive interviews; some students responded to the selected-response version of the argument by selecting all of the response options that they observed to be true based on the item scenario, including evidence that was true but tangential to their argument. In general, the selected-response format provided potential ideas for the students, and some students seemed to view their task as considering, and deciding whether to accept or reject them. On the other hand, the constructed-response versions of the items required students to generate their own ideas, a task which students tended to consider more difficult.

It is worth noting that students provided more information on selected-response versions of the items: more evidence and reasoning, and more answers overall. Some researchers have considered the selected-response item format as an additional type of scaffolding (Songer & Gotwals, 2012). The response options may have provided a point

of access for students who were confused by the task or unsure of how to answer a question, thus providing a form of scaffolding that enabled them to provide an answer where they otherwise couldn't. This is supported by data from the assessment pilot, where more "Missing on Engaging in Argument from Evidence" responses were observed when a constructed-response format was used, compared to a single-prompt format. From an assessment standpoint, the selected-response format may allow students to provide evidence of their understanding where there would otherwise be no evidence at all. This benefit must be weighed against the warping of the intended construct that occurs when students are allowed to consider ideas, rather than generate them.

Students were more likely to provide arguments on selected-response items. Response format did not appear to play as large of a role in eliciting relevant information about the three dimensions as multidimensional scaffolding. However, the use of a selected-response format increased students' likelihood of providing a response that included an argument compared to a constructed-response format. Creation of an argument from scratch is an effortful task, and the selected-response format removes much of that effort by providing response options. Reducing the effort required seemed to increase the likelihood of actually observing a response. However, it should be noted that selecting an appropriate argument from given options is a vastly different task than constructing an argument from scratch. The information gained by changing the response format may not outweigh the threat to validity posed by offering response options.

A selected-response format was associated with shorter response times. The selected-response item format was associated with shorter response times, on average, but the significance of the decrement is questionable: selected-response items averaged

only 28 fewer seconds of response time per item. Compared to the total average response time of ~3 minutes per item, selected response items decrease the response time by about 17%. Over multiple items, the differences in response time accumulate, suggesting that students would likely be able to complete more selected-response items than constructed-response items in a typical testing session. Thus, the decrease in response time associated with selected-response items could allow for a somewhat more reliable test that collected more information about student understanding. However, this consideration should be weighed against the effects of the selected-response item format on the intended construct.

The selected-response format was associated with more reliable scoring of arguments. The selected-response format was associated with higher interrater reliability for the argument dimension only, when a multidimensional scoring rubric was used. This finding held across all student ability subgroups. After scoring, many raters reported that the argument dimension was the most difficult to score, due to its complicated partial credit rubric and the wide range of student responses. The selected-response format allowed for a clearer rubric. Because students were limited to the given answer options, the scoring rubric was able to specify particular scores for different combinations of chosen responses, thereby providing extra clarity that was not possible in the constructed-response argument rubric. Therefore, it seems that the selected-response format facilitates clearer scoring rubrics by constraining student responses to a smaller number of possible variations which can be clearly specified.

The selected-response format did not have much impact on interrater reliability when a holistic rubric was used. It is unclear why – perhaps because the extra scaffolding

provided by the answer option was not compatible with the unstructured nature of a holistic rubric, or because the two non-argument dimensions produced a kind of "halo effect" whereby raters did not have to consider the more complex argument response as deeply as for the multidimensional rubric, thus cushioning their scores. Regardless, it seems that the selected-response format has the greatest impact on scoring reliability when the scoring rubric was specific and complex.

It was easier for students to generate arguments based on both evidence and reasoning when a selected-response format was used. On the Engaging in Argument from Evidence dimension, the highest item thresholds denote the point at which students move from generating arguments based on *either* evidence *or* reasoning, to arguments based on *both* evidence *and* reasoning. These highest thresholds tended to be easier to cross when a selected-response format was used. This suggests that the selected-response format may have changed the nature of the task in a way that made it substantially easier.

For tasks in which the student is asked to supply a scientific argument, a selected-response format may alter the intended construct by providing more support and suggestions. The creation of an argument is a complex practice, but the selected-response format provides a bit of extra scaffolding by providing answer options for students to consider. The answer options give students a framework for constructing an argument, as well as indicating potential sources of evidence and reasoning. The construct assessed by a selected-response argument item (i.e., ability to select from among several pieces of given evidence and reasoning) is substantially different from the construct assessed in a constructed-response item (i.e., ability to generate an original argument, drawing on the

student's own understanding as a basis for evidence and reasoning). This is a major concern with regard to assessment validity.

The selected-response format can improve or diminish item fit, depending on the task. In one case on the Structure and Properties of Matter dimension, the use of a selected-response format seemed to mitigate an issue that caused the open-ended version of the same item to have poor fit. Selected-response items provide extra scaffolding for students, which may help to narrow the scope of the task for students who are confused. In the item in question, unclear phrasing led to some confusion from the students. The response options mitigated the confusion, providing clarity that allowed students to give answers that better reflected their understanding and resulting in better item fit.

However, the selected-response format was associated with poor item fit on the Engaging and Argument from Evidence dimension. This finding was constant across all ability subgroups. The selected-response format added some additional challenges and supports to the students' testing experience. Compared to constructed-response argument items, the selected-response options required more reading from students. On the other hand, the response options also introduced evidence and reasoning that the students may not have otherwise considered. Both of these factors may have influenced students' responses in an atypical manner. When it comes to scientific argumentation, response options may change the way that students interact with the task, thereby increasing the amount of unpredictable variation in students' responses.

Reading load did not appear to impact item difficulty of selected-response argument items. Across all ability groups, there was no relationship between the difficulty of the lowest threshold values and response format, suggesting that the

increased reading load required by multiple-choice items does not disadvantage low ability students by providing an additional barrier to access. This finding contradicts conventional wisdom, which posits that selected-response items require extra reading comprehension which introduces construct-irrelevant variance. However, no information was available about the sample's overall socioeconomic status or English language proficiency, which may interact with reading demand.

Overall conclusions about response format. The selected-response format has some advantages but many more drawbacks, especially with regard to the Engaging in Argument from Evidence dimension. Many students report that they prefer the selected-response format because it helps them frame their response. Items that utilized a selected-response format required lesser response times, on average. Shorter response times allow more items to be administered within a testing period, and increases test reliability. A selected-response format may facilitate student access to more complex tasks, as students were more likely to provide any response to selected-response arguments than constructed-response arguments. Selected-response items limit the range of potential student responses, therefore allowing for easier, more reliable scoring. Response options provide a suggested response structure that may clarify the intended task, thereby increasing item fit. Contrary to expectation, the higher reading load of selected-response items did not cause a disadvantage to low-performing students.

However, the selected-response format also produced some threats to validity. The selected-response format seemed to add a certain form of construct-irrelevant variance relative to the constructed-response format, in that they introduce content in some student responses that may not have otherwise occurred. Despite expressing a

general preference for the selected-response format, students also reported being unsure of how to respond to selected-response items, especially on the Engaging in Argument from Evidence dimension – perhaps as a consequence of the amount and length of response options to argument items. To increase student comfort with the selected-response argument task, it is advisable that students be exposed to the format before encountering it on a test, perhaps as part of their classroom instruction. Furthermore, there is considerable evidence that the selected-response format resulted in a meaningful change to the nature of the argument task. Some students reported that the response options led them to rethink their answers. Students sometimes provided more detailed answers, or completely different answers when a selected-response format was used, compared to a constructed-response format. Providing arguments based on both evidence and reasoning was systematically easier when a selected-response format was used, based on a comparison of item difficulty estimates. These findings suggest that selected-response argument items may be measuring students' ability to recognize relevant/supporting evidence and/or reasoning, instead of the related, but more complex task of producing scientific arguments grounded in evidence and reasoning.

Overall, the selected-response format has different advantages and drawbacks for the assessment of scientific argumentation. In an assessment context, the choice of response format should be based on the inferences that assessment users want to make about students, considering how the different response formats might affect the strength of those inferences.

Research question 3: How do unidimensional and multidimensional scoring and modeling affect the empirical relationships among the 3 dimensions?

Model fit improves significantly with the multidimensional model. A direct comparison of model deviance statistics between the unidimensional and multidimensional approaches (both with a multidimensional/analytic scoring rubric) demonstrated that a multidimensional structure resulted in significantly better model fit than a unidimensional model structure. This indicates that a multidimensional structure better explains variation in students' responses to the assessment items than a unidimensional structure. This is also reflected in item fit statistics, which tended to improve under the multidimensional model – especially on the Engaging in Argument from Evidence dimension.

There are moderate-high correlations between dimensions. Based on the multidimensional model, the correlations between the 3 dimensions ranged from 0.73 to 0.89 – moderate to very strong correlations. The highest correlation was between the Scale, Proportion, and Quantity and Structure and Properties of Matter dimensions. This high correlation suggests that differentiation between the two dimensions may not be explaining any unique variance in student performance. The Engaging in Argument from Evidence dimension has somewhat lower correlations with the other two dimensions (0.73-0.80), suggesting that the Engaging in Argument from Evidence dimension is capturing some unique variation in student performance. A two-dimensional model where these two dimensions were collapsed fit the data significantly better than a unidimensional model, however, the three-dimensional model still demonstrated significantly better fit. However, the difference in deviance between the one- and two-

dimensional models was much larger than the difference between the two- and three-dimensional models, suggesting that specification of a separate Engaging in Argument from Evidence dimension has a much larger impact on model fit than the separation of Scale, Proportion, and Quantity and Structure and Properties of Matter. This may be due to the nature of these particular content areas, as the matter and measurement concepts tend to be heavily dependent on content-specific knowledge, while Engaging in Argument from Evidence is a more general skill.

The high correlation between dimensions may be bolstered by the fact that all test items are in the same narrow content area: matter, specifically, the matter concepts taught in Grade 4 of the Inquiry Project curriculum. If the crosscutting concept and science practice were tested in the context of other disciplinary core ideas, the correlation may not be so high.

An examination of disparities in WLE student ability estimates from the three-dimensional model demonstrates that about half of students have ability estimates that differ by a standard deviation or more on at least two dimensions, based on information from the current assessment. For these many students, the multidimensional estimates provide more detailed diagnostic information than would be gained from a unidimensional estimate. For example, teachers may utilize very different approaches for a student with a high Engaging in Argument from Evidence estimate, but a low Structure and Properties of Matter estimate compared to a student with moderate performance on both dimensions. The multidimensional ability estimates provide teachers with enough information to address students' diverse needs.

The relationship between dimensions fluctuates with ability. A subgroup analysis reveals that the relationship between dimensions may not be homoscedastic; specifically, the relationship between each pair of dimensions appears to be stronger among high-ability students than low-ability students. This indicates that the relationship between dimensions may be more complex than the overall correlations imply. When a student's mental model is sophisticated, their understanding of various matter and measurement concepts may coalesce, whereas a student's understanding may be more fragmented at lower levels. Once students have acquired some concepts, it may be easier for them to acquire other related concepts, strengthening the relationship between dimensions. For example, students who have a solid understanding of proportionality will be better equipped to pick up concepts about material density. Similarly, students who are familiar with evidence-based arguments may be better able to draw conclusions from data, thus enabling their acquisition of other concepts. On the other hand, students with partial or incomplete understanding may not experience the same benefits in acquisition of related concepts.

The heteroscedastic relationship between dimensions has implications for teaching and learning. It implies that students rely on already-acquired concepts as a foundation on which to build a more sophisticated understanding. Skills from one dimension should be referenced and utilized in the learning of the others, especially as students begin to master some of the concepts in a learning progression. This mirrors the NGSS emphasis that the three dimensions of science should be constantly integrated in teaching and learning.

Unidimensional reliability is higher than subscale reliability. Reliability is related to both the standard error of measurement and the number of items on a scale; therefore, it is not surprising to observe that both WLE and EAP reliability were higher when a unidimensional model was used to scale the data. Based on the EAP reliability estimate, reliability for the multidimensional subscales was slightly lower than unidimensional reliability (a drop of 0.07 to 0.12). However, based on the WLE reliability estimate, the drop in reliability was much more severe, especially on the Structure and Properties of Matter (difference in WLE reliability = 0.21) and Scale, Proportion, and Quantity dimensions (difference in WLE reliability = 0.40).

Multidimensional EAP person estimates take into account the relationship between dimensions, so these estimates tend to be more precise. Consequently, reliability estimates based on EAP estimates tend to be higher. On the other hand, multidimensional WLE estimates do not take the relationship between dimensions into account. They are unbiased estimates, but much more error-prone, and they lead to lower reliability estimates. It is therefore expected that WLE reliability will suffer more under the multidimensional approach, but the size of the decrement is shocking. The gap between EAP and WLE reliability indicates that it may be necessary to account for the correlation between dimensions in order to maintain adequate scale reliability under the multidimensional approach, a major concern when assessment is used for high-stakes decision making (like promotion, teacher accountability, or funding decisions). WLE estimates would certainly not be suitable for such purposes. On the other hand, EAP estimates, which reflect the high degree of correlation between dimensions, may not provide as much nuanced information as WLE estimates, which might be more

appropriate for lower-stakes decision making, like instructional adjustments for classrooms or individual students.

Due to the reliance of EAP estimates on the correlation between dimensions, EAP reliability also tends to fluctuate substantially depending on the strength of the relationship between dimensions in different sub-samples. Among low-ability students, the correlation between dimensions was much lower, and EAP reliability suffered. Among high-ability students, the correlations were stronger and reliability was greater. This illustrates how the nature of the relationship between dimensions is a critical consideration in a multidimensional assessment context, as it massively influences the quality and interpretability of assessment outcomes. When the relationship between dimensions is non-constant, the EAP estimates may overcorrect individual estimates to reflect the aggregate relationship between dimensions, inflating the reliability coefficient.

When a holistic rubric was used, both WLE and EAP reliability estimates from a unidimensional model were slightly lower than the multidimensional rubric (WLE reliability decreased by 0.03 from 0.82 under the multidimensional rubric to 0.79 under the holistic rubric, and EAP reliability decreased by 0.01 from 0.85 under the multidimensional rubric to 0.86 under the holistic rubric). The holistic model had a much smaller number of item thresholds, compared to the analytic models, and fewer items are generally associated with lower reliability. However in this case, the decrement in reliability is very small, suggesting that both models provide similar amounts of information despite the large discrepancy in number of item thresholds. This suggests that the additional score thresholds in the analytic models may not actually be adding precision to the ensuing measurements of student ability. This is a strong argument in

support of holistic rubrics, which tend to be easier for assessment developers to write and for raters to use; if they also provide estimates of student ability with just as much precision, it may not be worth going to the extra trouble to create more nuanced scoring rubrics.

Multidimensional person ability estimates have better fit for students with dissimilar performance on the separate dimensions. The multidimensional model had fewer students who underfit the model, meaning that there were fewer students whose responses were less predictable than the model assumed. This makes sense: assuming that multiple dimensions exist, students who have disparate performance on the different dimensions would have poor fit if a unidimensional model is used. A multidimensional model allows for discrepancies in student performance across the dimensions, leading to better person fit.

Multidimensional person ability estimates are less precise than unidimensional person ability estimates, but similar to holistic ability estimates. Person ability estimates from the unidimensional model tended to have smaller standard errors than multidimensional estimates, when the multidimensional rubric was used. On the Scale, Proportion, and Quantity dimension, especially, subscale estimates had much more uncertainty than the corresponding unidimensional estimates. The unidimensional model outdoes the multidimensional model when it comes to precise estimation of student ability. Unidimensional estimates tended to be more precise than the subscale estimates, when a multidimensional/analytic rubric was used. The size of person error is related to the number of items on a scale, so it follows that the smaller subscales would have larger errors than a larger unidimensional scale. Consequently, the Scale, Proportion, and

Quantity dimension, which had the smallest number of scale points, also had the largest standard error of measurement. The Structure and Properties of Matter and Engaging in Argument dimensions had a greater number of thresholds, and consequently they also had smaller increases in standard error of measurement relative to the unidimensional model.

Person estimates from the holistic rubric/unidimensional model had larger standard errors than the unidimensional estimates from the multidimensional model. Again, there were a smaller number of scale points when the holistic rubric was used, so the increase in error from the multidimensional rubric is expected. Standard error of person estimates were similar to the subscale person estimates, on average, when compared to scales with a similar number of thresholds (i.e., the Structure and Properties of Matter and Engaging in Argument from Evidence dimensions). The information functions (which are the inverse of square of the conditional standard errors) are therefore also similar. Even though the dimensional subscores focus on smaller aspects of the response, they still provide a similar amount of information as the holistic scores, which ostensibly contain a broader reflection of student performance.

Overall conclusions about dimensionality. It is impossible to say that one model structure or rubric categorization scheme is generally better or worse than another. Instead, each model has its own benefits and drawbacks, and depending on the intended use for student estimates a different model may be appropriate.

When the holistic rubric is used, reliability is satisfactory, and there is a good match between the distributions of item difficulty and person ability. Person fit and item fit are poor. A holistic rubric may not describe variations in student responses as well as

an analytic rubric. Raters may have difficulty implementing with fidelity a rubric based on broad categorizations of student responses, and this may lead to unpredictable patterns in scoring. These observations, combined with considerations about the broader violation of construct validity when multifaceted items are scored with a one-faceted holistic rubric, imply that the holistic scoring approach may not be appropriate.

When a multidimensional/analytic rubric is used to score student responses, the unidimensional model distributes item coverage evenly across a single scale. The unidimensional model also results in the most reliable scale, as it has the largest number of items. The correlation between dimensions is fairly high, lending credence to a unidimensional structure. However, additional evidence suggests that a three-dimensional structure does a better job of describing variation in student performance. Many students have large variations in performance across the dimensions. Furthermore, the multidimensional model results in slightly better item fit on the argument dimension, better person fit for students with disparate performance on the separate dimensions, and multidimensional model fit is significantly better than the unidimensional model. If model fit and nuanced information about student performance on each dimension are valued for the intended uses of test results, then the multidimensional model affords some advantages over the unidimensional model. Furthermore, the high dimensional correlations are based on the current assessment, which is limited to a single disciplinary core idea. Correlations may lessen when crosscutting concepts and science practices are assessed in the context of multiple disciplinary core ideas. If students' dimensional ability estimates diverge, multidimensional estimates would be even more informative.

The multidimensional model also exposes some large gaps in item coverage, suggesting areas of weakness that may have flown under the radar with a unidimensional model. On the Structure and Properties of Matter dimension, all thresholds are compressed in the center of the scale. The Engaging in Argument from Evidence thresholds are skewed towards the bottom of the scale. These gaps may lead to imprecise placement of those students with corresponding ability levels. Inserting new items that fill these gaps will improve measurement precision, regardless of the chosen dimensional structure.

The multidimensional model also results in subscale person estimates that have less precision than a unidimensional model, but similar to estimates based on a holistic rubric. Subscale reliability is lower than unidimensional reliability for both the holistic and multidimensional/analytic rubrics, especially when WLE person estimates are used to calculate the reliability estimate. This is not surprising, given that the subscales have a reduced number of items, relative to the unidimensional scale, but it should be taken into consideration, especially if reliability and precise estimation of person ability are high priorities of assessment. Using EAP person estimates instead of WLE person estimates mitigates the decrement in both the size of standard errors of measurement and in subscale reliability; however, EAP estimates should not be used if bias in person estimates is a concern. If high precision in student estimates is needed to, for example, evaluate whether a student has reached a certain threshold of performance based on a single cut point, then the combination of a unidimensional model with a multidimensional rubric seems best suited to this task. On the other hand, subscale estimates are most useful in instructional settings, where high reliability is not as critical.

Finally, it should be noted that subgroup analysis demonstrated that the relationship between dimensions may not be constant across students of all abilities. Further research should be done to tease apart the relationship among the three NGSS constructs studied here: Structure and Properties of Matter, Scale, Proportion, and Quantity, and Engaging in Argument from Evidence, as well as the relationships between these constructs and other related science constructs. Gaining more information about the relationships between dimensions in many contexts and content areas will contribute to continued improvement in the understanding of science as a multidimensional construct, which will further lead to new strategies for teaching and learning science and better measurement of science at every age.

Research question 4: Instrument Validity

Gaps in item difficulty are masked when a unidimensional model is used. On the Structure and Properties of Matter and Engaging in Argument from Evidence dimensions, item thresholds did not span the entire dimensional subscale, leaving entire segments of the scale without adequate item coverage. On the Structure and Properties of Matter dimension, only the middle of the scale contained adequate item coverage. On the Engaging in Argument from Evidence dimension most of the item thresholds fell near or below the scale mean, revealing an upper region of the scale that was only sparsely populated with items. This exposes the need for more items to fill in the gaps on the Structure and Properties of Matter and Engaging in Argument from Evidence dimensions. The Scale, Proportion, and Quantity scale had adequate item coverage across the entire scale range. When a unidimensional model was used, there were no visible gaps in

coverage, masking the failure of the item pool to adequately measure the entire range of student abilities on specific subdomains.

Discrepancies between hypothesized and observed item difficulty distributions suggest areas for revision of the construct definition. On the Structure and Properties of Matter dimension, it was easier than anticipated for students to recognize weight invariance in the presence of a phase change. This suggests that once students understand the basic principle of conservation of weight, the leap from physical to chemical changes may not present as great an obstacle as previously anticipated.

Furthermore, it was much harder for students to make overall judgments about whether objects could be made of the same material when both the weight and the volume were different, compared to items where either the weight or the volume was held constant. This is a factor that should be accounted for in the construct map. The item in question also ties in topics from the Scale, Proportion, and Quantity construct, like proportionality. This concept may conflate the two constructs such that they are dependent upon each other, which may have increased item difficulty.

On the Scale, Proportion, and Quantity dimension, items that required multiplication and division tended to be more difficult than items that involved addition and subtraction. The complexity of the mathematical calculation seems to be an important factor, perhaps an auxiliary construct that should be recognized and included as part of the operational definition for this dimension.

Otherwise, the anticipated order of item difficulty largely aligned with the observed estimates. The three sub-constructs appear to be well-defined, and the items seem to be a valid representation of those constructs. Some small revisions to the

construct maps may improve the alignment between hypothesized and observed item difficulty.

Gaps in item coverage should be filled. If a multidimensional approach is utilized in the future, effort should be made to fill in the gaps on the Structure and Properties of Matter and Engaging in Argument from Evidence scales, both of which display large gaps in item difficulty. On the Structure and Properties of Matter dimension, most item thresholds are clustered near the center of the scale. Since all scale items were based on concepts from one year (grade 4) of the Inquiry Project Curriculum, it makes sense that the concepts measured by the items would have similar difficulty. To increase the range of item thresholds, concepts from the earlier and later grades could be added. If done in conjunction with the assessment of higher and lower-grade concepts from the Inquiry Project, this would also enable the construction of a vertical scale, in which items from the third, fourth, and fifth grade curriculum could be placed on a common scale and student growth could be measured across grades.

On the Engaging in Argument from Evidence dimension, most item thresholds are clustered near the bottom of the scale. This leaves the upper regions open, without adequate items to measure students with higher ability on this dimension. This also affects the standard error of person estimates on the argument dimension, resulting in higher error rates near the mean of the person distribution than are usually observed. The definition of the Engaging in Argument from Evidence construct may need to be revised so that it better describes above-average student performance, and the items/rubrics rewritten to capture variation at the high end of the scale.

Item fit statistics suggest items that could be replaced or improved. Some item fit issues are likely related to response format (see KEVIN), however, there are some additional items with a high degree of misfit. These items should be examined by content and curriculum experts to ensure that student responses are not influenced by construct irrelevant factors. It would also be useful to look at patterns in individual student residuals, as they may provide more detailed information about the types of responses that tended to result in misfit. Based on findings, these items should be considered as candidates for revision or replacement.

It is likely that rater effects have contributed to some poor fit on the Engaging in Argument dimension. As stated in previous sections, revisiting the scoring rubric for this dimension is a high priority to improve scoring reliability. It may also help improve the poor item fit.

Item performance gaps between Inquiry Project students and non-Inquiry students suggest strengths and shortcomings of the Inquiry Project curriculum. A DIF analysis revealed that fifth grade *Inquiry Project* students performed unexpectedly poorly on items assessing the invariance of volume of solid objects when reshaped, when compared to non-*Inquiry* students of similar ability; however, *Inquiry* students did better than their equal-ability non-*Inquiry* counterparts on items assessing whether tiny objects have weight and volume. Both concepts are covered extensively in the *Inquiry Project* curriculum, and the pattern suggests that participation in the *Inquiry Project* curriculum may affect the way that students interact with these items. For items 8A and 10, which assess the weight and volume of tiny objects, the curriculum may have improved student performance. However, the curriculum may have detracted from student performance on

items 1A, 1B, and 11, which ask students to answer whether the volume of a rectangular solid changes after being rearranged. This concept is discussed in the *Inquiry Project* curriculum at length; in fact, an entire lesson is devoted to a near-exact scenario as that in items 1A/1B, yet *Inquiry Project* students are performing worse than their non-*Inquiry* counterparts on this item. The item should be re-examined, to make sure that the language is consistent with the way it was taught in the curriculum, and the curriculum should be reexamined to make sure that the lessons are productive for students and not unintentionally reinforcing misconceptions.

No other clear patterns in DIF emerge. Although DIF is observed on a few more items when student performance is examined by grade and by curriculum participation, there are no clear item or subgroup characteristics that may account for differential performance. It is possible that in these cases, DIF was observed by chance due to small subgroup sample sizes[40] or unreliability in scoring[41], or that it was caused by some other factor unknown to the researcher. These limitations should be addressed in the future.

Inquiry Project students score higher than non-Inquiry Project students on all three dimensions. Average student performance is higher among *Inquiry Project* students on all three dimensions. The most likely explanation for this pattern is that curriculum participation gave students an advantage on the assessment. Given that the assessment was written to align with the *Inquiry Project* curriculum, it makes sense that curriculum participants should have an advantage over students who did not participate in the curriculum. However, there are other background factors that were associated with

[40] Sample sizes ranged from N=17 to N=197, depending on subgroup classification and test form exposure, with category frequencies as low as N=1.
[41] This was especially concerning on the Engaging in Argument from Evidence dimension. For a discussion of scoring reliability, see Chapter 4, pg 155.

Inquiry Project participation, including location (district and state). Therefore, it is possible that average differences in student performance are due, at least in part, to systematic differences in these other factors. Further research should use matched comparison groups of *Inquiry Project* and non-*Inquiry* students to control these potential confounding explanatory variables and better isolate the effect of the curriculum.

Fifth-grade students score higher than fourth-grade students on all three dimensions. Among students who participated in the *Inquiry Project* curriculum, 5th grade students had higher average performance across all three dimensions than did 4th grade students. The difference between grades ranged from about 20% to 40% of a standard deviation, depending on the subscale. This finding conforms with the expectation that students will be better able to demonstrate understanding of the concepts covered in the assessment after an additional year of exposure to the curriculum. Furthermore, since the assessment was designed to cover concepts taught in the 4th grade, it is plausible that performance will be higher among students who completed the 4th grade curriculum (in this study, the 4th grade students had not yet completed the 4th grade curriculum).

Within-classroom variability is larger among Inquiry Project students on the Engaging in Argument from Evidence dimension. Regardless of grade level, within-classroom variability of student performance on the Engaging in Argument from Evidence dimension tended to be larger in classrooms that participated in the *Inquiry Project* curriculum, compared to classrooms that did not participate. This suggests that student understanding of Engaging in Argument from Evidence becomes more spread out after classrooms participated in the curriculum. One potential explanation for this finding

is that the *Inquiry Project* curriculum successfully enhances student proficiency with scientific argumentation, but has more impact on students who already have a high ability. This would create more separation between the highest and lowest ability students, increasing within-classroom variability among curriculum participants. Were this the case, more work should be done to attend to lower-ability students in *Inquiry Project* classrooms, so that they may enjoy the same benefits in learning as their high-ability classmates. Again, this finding is limited by the presence of other factors that were associated with *Inquiry Project* participation, including location (district and state). To eliminate the influence of confounding local variables, further research should examine within-classroom variability among matched comparison groups of *Inquiry Project* and non-*Inquiry* students.

There are no differences in within-classroom variability among *Inquiry Project* and non-*Inquiry Project* participants on the remaining two dimensions, or between grade 5 and grade 4 classrooms on any dimension. This suggest that student understanding of the Scale, Proportion, and Quantity and Structure and Properties of Matter dimensions does not become more uniform with age or exposure to the curriculum.

Between-classroom variability is high among Grade 5 Inquiry Project classrooms on the Structure and Properties of Matter dimension. Among Grade 5 students who participated in the *Inquiry Project* curriculum, 19% of variability in student performance on the Structure and Properties of Matter dimension is attributable to classroom membership. This finding suggests that student performance may have been affected by differences in how individual teachers taught the curriculum, or other classroom-specific factors. Among 4th grade *Inquiry Project* participants and 5th grade

non-*Inquiry* participants, a much smaller percentage of between-classroom variability was observed, suggesting that these groups' performances were not as affected by classroom-specific factors.

Between-classroom variability is high among Grade 5 non-Inquiry Project classrooms on the Engaging in Argument from Evidence dimension. Among Grade 5 students who did not participate in the *Inquiry Project* curriculum, 18% of variability in student performance on the Engaging in Argument from Evidence dimension is attributable to classroom membership, suggesting that differences in teaching among different classrooms/schools may have affected student performance. Teaching and learning about scientific argumentation is much more uniform among classrooms that participate in the *Inquiry Project* curriculum, compared to classrooms that do not participate. The presence of an inquiry-based curriculum may even out imbalances in individual teachers' focus on scientific argumentation, accounting for the low amount of between-classroom variability among *Inquiry Project* classrooms. However, this finding could again be attributable to other local factors, as location was confounded with *Inquiry Project* participation in the study design.

Overall conclusions. Assessment results suggest some small discrepancies between observed patterns of student performance and the hypothesized progression of student understanding based on the *Inquiry Project* curriculum, but overall the results align with the underlying structure of the curriculum. There are two potential areas for improvement to the *Inquiry Project* curriculum, most notably, that *Inquiry Project* students seem to struggle with volume invariance of solid objects when reshaped. This

suggests that instruction related to this concept may be inadvertently introducing a misconception.

As anticipated, *Inquiry Project* students and 5th grade students tended to demonstrate better understanding of the assessed concepts. This was true of all three assessment dimensions. However, local factors may confound the comparison of *Inquiry Project* and non-*Inquiry* performance, so the difference cannot be attributed solely to *Inquiry Project* participation. Future studies should use matched comparison groups to eliminate confounding background variables and isolate the effect of the *Inquiry Project* curriculum on student understanding.

On the Engaging in Argument from Evidence dimension, there appears to be more variability in student performance within *Inquiry Project* classrooms than within non-*Inquiry* classrooms. Student performance is higher, on average, in *Inquiry Project* classrooms compared to non-*Inquiry* classrooms, by about a third of a standard deviation. If these differences are attributable to the curriculum, it suggests that the curriculum may help students who already have above average skill in scientific argumentation more than low-ability students, increasing the range of student performance within each classroom. If this is the case, future work should explore best practices for teaching scientific argumentation to low-ability students.

Additionally, there exist larger differences between classrooms in some subgroups, depending on the particular assessment dimension. The grade 5 *Inquiry Project* sample demonstrated substantively larger between-group variability than non-*Inquiry* students on the Structure and Properties of Matter dimension (17.5% of variability exists between grade 5 Inquiry Project classrooms, compared to less than 4%

of variability between non-Inquiry classrooms), and the grade 5 non-*Inquiry Project* students demonstrated substantively larger between-group variability on the Engaging in Argument from Evidence dimension (11.8% of variability exists between grade 5 non-Inquiry classrooms, compared to less than 6% of variability between Inquiry Project classrooms). In both cases, school or teacher differences may affect student performance.

Implications

The implications for the findings of this study depend on the intended usage of assessment results. Therefore, three separate groups of users will be considered, and varying implications presented for each:

1. Teachers, and other school and district science support staff, who gather information about student understanding to influence instructional decision-making.
2. State and district policymakers, who gather student performance data for monitoring and accountability purposes.
3. Researchers and evaluators, who gather evidence of student understanding to look for growth associated with a particular curriculum or educational intervention.

Implications for teachers and science support staff. When it comes to item structure, it is clear that assessment scaffolding enriches the amount and quality of evaluable evidence provided by students on complex assessment tasks, especially among lower ability students. Scaffolding, therefore, should be a valuable tool for teachers who are trying to collect information about multidimensional student ability. However, it may also be useful for teachers to augment scaffolded tasks with a few broader, open-ended tasks without scaffolding, as these can be an indicator of how well students see the bigger

picture and are able to manage and organize their understanding of the dimensions in a cohesive way.

With regard to response format, the findings showed that multiple-choice response options tend to introduce reasoning that students may not have otherwise considered. This was especially true in the domain of scientific argumentation. However, students seemed to enjoy the selected-response items, and some students suggested that the response options could be used as a basis for constructing a written argument. Perhaps selected-response arguments could be a useful instructional tool and supporting guide for students who are still mastering the concepts and learning how to construct a scientific argument.

When it comes to scoring and scaling, deliberation about whether or not dimensional subscores are distinct enough may not be practically useful to classroom teachers. For teachers, tracking student performance is a way to monitor and guide instruction. Since the NGSS emphasizes multidimensional science instruction, tracking students' progression on all aspects of science is important. Multidimensional/analytic scoring rubrics are a useful tool to help teachers meet this goal, as they provide important structure to gather information about student understanding on all relevant aspects of the instructional content. Multidimensional rubrics also facilitate more information about students' failure to provide a response that addresses all relevant dimensions, an important indicator that students struggle to attend to and organize responses that reflect multiple dimensions of understanding.

Implications for state and district policymakers. For monitoring and accountability purposes, high-quality, reliable information is of paramount importance.

Scaffolding affords more thorough information about multidimensional student understanding, and therefore seems a prudent choice for item design, especially for students in lower grades (i.e., those with lower abilities and those who struggle to integrate multiple aspects of learning/instruction into a complex response). However, assessment design has a large impact on the direction of instruction, and inclusion of more open-ended tasks in small amounts or at higher grade levels may be a necessary to encourage some instructional focus on "the big picture," i.e., developing students' ability to draw together knowledge/skills from different dimensions in a cohesive manner.

The multiple-choice response format is useful for large-scale assessment because it allows for a large number of items to be administered in a shorter timeframe, and is less costly to score. This makes multiple-choice items very desirable for assessment. However, the findings here suggest that the multiple-choice format may not be suitable for all dimensions of science learning; in particular, the Engaging in Argument from Evidence dimension was not well-represented by multiple-choice items. The selected-response format seemed to change the nature of the task by prompting students to consider new ideas that may not have otherwise, and by encouraging students to select multiple responses, they may have been encouraged to choose what seemed true based on the assessment prompt, rather than what was the best supporting evidence/reasoning. This indicates a discrepancy between the intended and enacted construct. Altogether, a selected-response format may be an efficient and valid format for assessing some dimensions, but its utility for measuring science practices (in particular, scientific argumentation) is questionable.

With the exception of the argumentation dimension, multidimensional (analytic) scoring tended to have higher interrater reliability than holistic scoring, especially when used to evaluate student responses to highly scaffolded items. Depending on the extent of usage of the selected-response format in subtasks, multidimensional scoring may also lend itself to a higher degree of automated scoring, while holistic scores require rater judgment. Therefore, with the paramount consideration of reliability and cost, multidimensional scoring seems to be the obvious choice, with the caveat that scoring of scientific practices (e.g., argumentation) should be considered carefully, as rater judgments appear to be very difficult on this dimension.

In the context of the current example, the utility of multidimensional subscores as information for monitoring and accountability is highly suspect. Accountability estimates like teacher value-added or student growth percentiles carry high stakes, and measurement error affects the quality of these estimates (National Research Council, 2011). Therefore, error in measurement should be minimized and reliability maximized. When WLE person estimates are used, subscale reliability is insufficient for high stakes decisions about teachers, students, or schools. When EAP person estimates are used, most of the nuance in student performance across dimensions is absorbed by the high correlation between dimensions, rendering the subscale estimates largely uninformative. Based on the current study, a unidimensional scale is far more reliable, with only a small loss of information about student performance. However, it should be noted that the current study took place in a very narrow content domain, and that correlations between dimensions may have been strengthened because of this. It is unlikely that an assessment with such a small scope would be used for large-scale monitoring and accountability;

therefore, results may not generalize to real world situations. More research should be done to establish the correlation between dimensions on assessments with broader content domains.

Implications for researchers and evaluators. To examine the impact of the *Inquiry Project* curriculum and other similar educational interventions, it is important that assessment reflects the core objectives of the curriculum. In this case, since the curriculum incorporated elements of several NGSS dimensions, a multidimensional assessment was necessary in order to reflect the content of the curriculum. Item scaffolding was useful tool in support of multidimensional assessment, as it aided in gathering more, higher quality evidence about student understanding linked back to the curriculum. Scaffolding may be useful in other research contexts where gathering information about student understanding of complex or multidimensional constructs is a goal.

With regard to response format, the appropriateness of constructed- or selected-response depends on the construct being measured for change. In this case, constructed-response items more authentically captured student argumentation ability. It is possible that the constructed-response format would also be more appropriate for other practices like Constructing Explanations, in which originality is an important aspect of the construct. For more content-related constructs, a multiple-choice format may be appropriate.

When researchers and evaluators look for changes in student learning, measurement error is a major concern. Unreliable assessment data adds noise to the estimation of the effect of an intervention. Based on the lack of reliability of the

multidimensional subscales, a unidimensional scale will probably provide more precise estimates of overall change in student understanding. However, if the curriculum/intervention's purpose is to effect multidimensional change, then a unidimensional estimate will not provide sufficient information about the intervention effect. In the current study, differences between groups in student performance on the Scale, Proportion, and Quantity, Structure and Properties of Matter, and Engaging in Argument from Evidence dimensions were observed, even though the individual subscales had low reliability estimates. In similar studies, concerns about measurement error should be weighed against the need for valid representation of the construct(s) of interest.

Limitations

This study was limited to only one combination of NGSS constructs. The NGSS describe hundreds of other combinations in performance expectations, which involve the integration of at least one Disciplinary Core Idea, Crosscutting Concept, and Science and Engineering Practice. Thus, results may not generalize to other NGSS constructs, or multidimensional constructs.

This study was limited to 4th and 5th graders, so findings may not be generalizable to students of other ages. In particular, this study showed that the effect of scaffolding varied depending on student ability, suggesting that results may vary for a sample of more or less proficient students. Scaffolding had the smallest effect on high ability students, so it may be true that the effect of scaffolding may become less pronounced as students develop a more sophisticated understanding of science concepts. Additionally, the strength of the relationship among the three assessment constructs may vary at

different ages. In elementary grades, inquiry is a natural way that students learn about the world, so the relationship between content and practice may be stronger than it is at older grades.

When drawing comparisons among different scaffolding variations and response formats, the analysis was limited by the small number of potential item comparisons; there were only four response format pairs, and five scaffolding pairs, which made it difficult to make generalizable statements about the impact of response format and scaffolding on item fit, difficulty, interrater reliability and other psychometric criteria.

Differences in rater judgment limited the extent to which comparisons could be made between argument prompts with different response formats and levels of scaffolding. Although a rater model was employed to account for differences between raters on this dimension, raters and items were not fully crossed, leaving substantial room for uncertainty in the estimates of rater effects. Consequently, comparisons drawn between items and scores on the Engaging in Argument from Evidence dimension may not capture the true effect of those item variations and may not be generalizable to all raters.

Finally, the study design limited the extent to which comparisons could be made between Inquiry and non-Inquiry students due to the presence of confounding contextual factors, so the isolated impact of the curriculum on student understanding remains to be seen.

Directions for future work

Based on the findings of this study, there is room for improvement in some features of the assessment. Analysis revealed that there were gaps in the Structure and

Properties of Matter subscale, such that there were few items with above and below average difficulty. This resulted in less precise measurement for high and low ability students. Additional items should be added to fill in these gaps, perhaps by including material from the 3rd and 5th grade *Inquiry Project* curriculum. This would have the added benefit of facilitating vertical scaling for any future assessment of 3rd and 5th grade *Inquiry Project* concepts.

Additionally, the interrater reliability analysis revealed substantial shortcomings of the Engaging in Argument from Evidence scoring rubric. Sources of rater discrepancies should be examined, by engaging in a qualitative exploration of any common features of responses that were given discrepant scores, and/or by consultation with raters to get their firsthand account of their experiences using the rubric.

To get a better sense of the effect of the Inquiry Project curriculum on student understanding, future research should use matched comparison groups to isolate the effect of the curriculum while accounting for other background factors that may be related to student performance. Additional data should be collected on other factors of interest (i.e., student gender, race, and SES), to provide insight on the curriculum's effectiveness among different subgroups.

Further research should also examine the impact of raters on the generalizability of scores on constructs that require complex judgment (such as scientific argumentation). This analysis was limited by the use of a nested design; further research should use a more robust fully crossed design to draw stronger conclusions about rater effects, and work towards creating rubrics/scoring guides that minimize rater effects. For future work with multidimensional scoring, it may be useful to examine rater cognition through

interviews and cognitive labs, to determine whether the presence of multiple dimensions creates a "halo effect", prompting raters to make unfounded judgments of student performance. If this effect exists, it would be interesting to determine whether or not the effect can be mitigated by features of the rubric or item/response structure.

Research should be extended to students of different ages, especially when it comes to a) the effect of scaffolding on student responses, and b) the relationship between different dimensions of science learning. This study produced tentative evidence that the relationship between dimensions appears to lessen at the highest levels of student ability, which suggests that the relationships among the three dimension may not be as strong for older students.

Difference in the strengths of the relationships may impact the choice of scaling and reporting approach for students of different ages. As for scaffolding, it may be possible to get a greater amount of evaluable information about the understanding of older students with less prompting. Once students develop more sophisticated understanding of science concepts, fading the scaffolding would allow for more face validity in the assessment of an integrated multidimensional construct.

Finally, the work done here should be replicated among additional NGSS constructs. The NGSS performance expectations provide a multitude of other examples of settings in which the three dimensions can be combined for a particular topic and grade level, and these can be used as the basis for additional assessments. Some effects may vary depending on the specific combination of NGSS topics, including the effects of item structure and response format, and the relationships among dimensions.

References

Ackerman, T. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*(1), 67-91.

ACT. (2016). *Description of the ACT.* Retrieved from http://www.actstudent.org/testprep/descriptions/?_ga=1.175192930.1757403675.1454189983

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*(1), 1-23.

Adams, R. J., Wu, M. L., & Wilson, M. (2015). ACER ConQuest 4 [computer software]. Camberwell, Australia: Australian Council for Educational Research.

Agustin, T. (2006). An adjustment for sample size in DIF analysis. *Rasch Measurement Transactions,* 20(3), 1070-1071.

Almond, P. J., Cameto, R., Johnstone, C. J., Laitusis, C., Lazarus, S., Nagle, K., Parker, C. E., Roach, A. T., & Sato, E. (2009). *White paper: Cognitive interview methods in reading test design and development for alternate assessments based on modified academic achievement standards (AA-MAS).* Dover, NH: Measured Progress and Menlo Park, CA: SRI International.

American Association for the Advancement of Science. (1993). *Benchmarks for scientific literacy.* New York, NY: Oxford University Press.

American Association for the Advancement of Science. (2001). *Atlas of scientific literacy.* Washington, DC: AAAS.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology, 37*, 1-15.

Barrow, L. H. (2006). A brief history of inquiry: From Dewey to standards. *Journal of Science Teacher Education, 17*, 265-278.

Baxter, G. P., Shavelson, R. J., Herman, S. J., Brown, K. A., & Valadez, J. R. (1993). Mathematics performance assessment: Technical quality and diverse student impact. *Journal for Research in Mathematics Education, 24*(3), 190-216.

Beatty, P. C., & Willis, G. B. (2007). *Research synthesis: The practice of cognitive interviewing. Public Opinion Quarterly, 71*(2), 287-311.

Berg, C. A., & Smith, P. (1994). Assessing students' abilities to construct and interpret line graphs: Disparities between multiple-choice and free-response instruments. *Science Education, 78*(6), 527-554.

Berland, L. K., & McNeill, K. L. (2010). A learning progression for scientific argumentation: Understanding student work and designing supportive instructional contexts. *Science Education, 94*(5)765-793.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy.* New York, NY: Academic Press.

Boone, W. J., & Scantlebury, K. (2006). The role of Rasch analysis when conducting science education research utilizing multiple-choice tests. *Science Education, 90*(2), 253-269.

Braun, H. I., & Mislevy, R. (2005). Intuitive test theory. *The Phi Delta Kappan, 86*(7), 488-497.

Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment, 11*(1), 33-63.

Briggs, D. C., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement, 4*(1), 87-100.

Brown, N. J. S., & Wilson, M. (2011). A model of cognition: The missing cornerstone of assessment. *Educational Psychology Review, 23*, 221-234.

Carnegie Corporation of New York. (2009). *The Opportunity Equation: Transforming Mathematics and Science Education for Citizenship and the Global Economy.* New York, NY.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29*(1), 3-13.

Chiappetta, E. L., & Fillman, D. A. (2007). Analysis of five high school biology textbooks used in the United States for inclusion of the nature of science. *International Journal of Science Education, 29*(15), 1847-1868.

Chin, C., & Teou, L. (2009). Using concept cartoons in formative assessment: Scaffolding students' argumentation. *International Journal of Science Education, 31*(10), 1307-1332.

The College Board. (2014a). *AP Chemistry Course and Exam Description: Revised Edition.* New York, NY: The College Board.

The College Board. (2014b). *AP Physics 1: Algebra-Based and AP Physics 2: Algebra-Based Course and Exam Description Including the Curriculum Framework: Revised Edition.* New York, NY: The College Board.

The College Board. (2014c). *Getting Ready for the SAT.* New York, NY: The College Board.

The College Board. (2015). *AP Biology Course and Exam Description: Revised Edition.* New York, NY: The College Board.

Committee on a Conceptual Framework for New K-12 Science Education Standards. (2011). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas.* Washington, DC: The National Academies Press.

Davey, T., & Hirsch, T. M. (1991). Concurrent and consecutive estimates of examinee ability profiles. Paper presented at the *Annual Meeting of the Psychometric Society,* New Brunswick, NJ.

de Ayala, R. J. (2009). *The theory and practice of item response theory.* New York, NY: The Guilford Press.

deBarger, A. H., Harris, C. J., D'Angelo, C., Krajcik, J., Dahsah, C., Lee, J., & Beauvineau, Y. (2014). Constructing assessment items that blend core ideas and science practices. In Polman, J. L., Kyza, E. A., O'Neill, D. K., Tabak, I., Penuel, W. R., Jurow, A. S., O'Connor, K., Lee, T., & D'Amico, L. (Eds.). *Learning and Becoming in Practice: The International Conference of the Learning Sciences (ICLS) 2014, Volume 3.* Boulder, CO: International Society of the Learning Sciences.

Drennan, J. (2003). Cognitive interviewing: Verbal data in the design and pretesting of questionnaires. *Journal of Advanced Nursing, 42*(1), 57-63.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association, 56*(293), 52-64.

Duschl, R. A., Schweingruber, H. A., & Shouse, A. W. (2007). *Taking Science to School: Learning and Teaching Science in Grades K-8.* Washington, D. C.: The National Academies Press.

Dwyer, A., Boughton, K. A., Yao, L., Steffen, M., & Lewis, D. (2006). A comparison of subscale score augmentation methods using empirical data. Paper presented at the *Annual Meeting of the National Council on Measurement in Education*, San Francisco, CA.

Eltinge, E. M., & Roberts, C. W. (1993). Linguistic content analysis: A method to measure science as inquiry in textbooks. *Journal of Research in Science Teaching, 30*(1), 65-83.

Enright, M. K., Morley, M., & Sheehan, K. M. (2002). Items by design: The impact of systematic feature variation on item statistical characteristics. *Applied Measurement in Education, 15*(1), 49-74.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data.* Cambridge, MA: MIT Press.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika, 10*(4), 507-521.

Florida Department of Education. (2015). *Next generation sunshine state standards.* Retrieved from: http://www.cpalms.org/Public/search/Search

Folk, V. G., & Green, B. F. (1989). Adaptive estimation when the unidimensionality assumption of IRT is violated. *Applied Psychological Measurement, 13*(4), 373-389.

Fraser, C., & McDonald, R. P. (2003). NOHARM: A Windows program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory [Computer program]. Welland, ON: Niagara College. Available at www.niagarac.on.ca/~cfraser/download/.

Gorin, J. S. (2006). Test design with cognition in mind. *Educational Measurement: Issues and Practice, 25*(4), 21-35.

Gorin, J. S., & Mislevy, R. J. (2013). Inherent measurement challenges in the Next Generation Science Standards for both formative and summative assessment. Proceedings from the *Invitational Research Symposium on Science Assessment*. Princeton, NJ: K-12 Center at ETS.

Gotwals, A. (2006). *Students' science knowledge bases: Using assessment to paint a picture* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (3237965)

Gotwals, A.W., & Songer, N.B. (2006a). Measuring students' scientific content and inquiry reasoning. *Proceedings of the 7th International Conference of the Learning Sciences.* Bloomington, IN

Gotwals, A. W., & Songer, N. B. (2006b). *Cognitive predictions: BioKIDS implementation of the PADI assessment system (PADI Technical Report 10).* Menlo Park: SRI International.

Gotwals, A. W., & Songer, N. B. (2010). Reasoning up and down a food chain: Using an assessment framework to investigate students' middle knowledge. *Science Education, 94*, 259-281.

Gotwals, A. W., & Songer, N. B. (2013). Validity evidence for learning progression-based assessment items that fuse core disciplinary ideas and science practices. *Journal of Research in Science Teaching, 50*(5), 597-626.

Gotwals, A., Songer, N. B., & Bullard, L. (2012). Assessing students' progressing abilities to construct scientific explanations. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science: Current challenges and future directions.* Rotterdam, the Netherlands: Sense Publishing.

Gronmo, L. S., Lindquist, M., Arora, A., & Mullis, I. V. S. (2013). TIMSS 2015 mathematics framework. In Mullis, I. V. S., & Martin, M. O. (Eds.), *TIMSS 2015 Assessment Frameworks.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Harrell, L. M., & Wolfe, E. W. (2009). *Effect of between-dimension correlation and sample size on multidimensional Rasch analysis.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher, 30*, 141-158.

Howell, H., Phelps, G., Croft, A. J., Kirui, D., & Gitomer D. (2013). *Cognitive interviews as a tool for investigating the validity of content knowledge for teaching assessments* (Research Report RR-13-19). Princeton, NJ: Educational Testing Service.

Illinois State Board of Education. (2013). Test maps for the 2014 Illinois Standards Achievement Test. Retrieved from: http://www.sps186.org/downloads/basic/379362/ISAT%202014%20Blueprint.rtf

Jones, L. R., Wheeler, G. & Centurino, V. A. S. (2013). TIMSS 2015 science framework. In Mullis, I. V. S., & Martin, M. O. (Eds.), *TIMSS 2015 Assessment Frameworks.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Kane, M. T. (2006). Validation. In Brennan, R. L. (Ed.), *Educational Measurement* (17-64). Westport, CT: Praeger Publishers.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50(1), 1-73.

Kang, H., Thompson, J., & Windschitl, M. (2014). Creating opportunities for students to show what they know: the role of scaffolding in assessment tasks. *Science Education, 98*, 674-704.

Katz, S., & Lautenschlager, G. L., (2001). The contribution of passage and no-passage factors to item performance on the SAT reading task. *Educational Assessment, 7*(2), 165-176.

Kennedy, C. A., & Wilson, M. (2007). Using progress variables to interpret student achievement and progress (BEAR Technical Report No. 2006-12-01). Berkelely, CA: Berkeley Evaluation and Assessment Research Center.

Kirschner, P. A., & Meester, M. A. M. (1988). The laboratory in higher science education: Problems, premises and objectives. *Higher Education, 17*, 81-98.

Knafl, K., Deatrick, J., Gallo, A., Holcombe, G., Bakitas, M., Dixon, J., & Grey, M. (2007). Focus on research methods: The analysis and interpretation of cognitive interviews for instrument development. *Research in Nursing and Health, 30*, 224-234.

Kuusela, H., & Paul, P. (2000). A comparison of concurrent and retrospective verbal protocol analysis. *American Journal of Psychology*, *113*(3), 387–404.

Lau, K. (2009). A critical examination of PISA's assessment on scientific literacy. *International Journal of Science and Mathematics Education, 7*, 1061-1088.

Lee, H., Liu, O. L., & Linn, M. C. (2011). Validating measurement of knowledge integration in science using multiple-choice and explanation items. *Applied Measurement in Education, 24*, 115-136.

Le Hebel, F. Montpied, P. & Tiberghien, A. (2014). Which effective competencies do students use in PISA assessment of scientific literacy? In C. Bruguiere, A. Tiberghien, & P. Clement (Eds.), *Topics and Trends in Current Science Education: 9th ESERA Conference Selected Contributions* (p. 273-289). New York, NY: Springer.

Linacre, J. M. (1989). *Many-faceted Rasch measurement.* Chicago: MESA Press.

Linacre, J. M. (1994) Sample size and item calibrations stability. *Rasch Measurement Transactions, 7*(4), 328.

Linn, R. L., & Herman, J. L. (1997). *Standards-led assessment: Technical and policy issues in measuring school and student progress*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for the Study of Evaluation (CSE), Graduate School of Education & Information Studies, University of California, Los Angeles.

Liu, O. L., Lee, H., & Linn, M. C. (2011). An investigation of explanation multiple-choice items in science assessment. *Educational Assessment, 16*, 164-184.

Liu, O. L., Wilson, M., & Paek, I. (2008). A multidimensional Rasch analysis of gender differences in PISA mathematics. *Journal of Applied Measurement, 9*(1), 18-35.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Ludlow, L. H., Matz-Costa, C., Johnson, C., Brown, M., Besen, E., & James, J. B. (2014). Measuring engagement in later life activities: Rasch-based scenario scales for work, caregiving, informal helping, and volunteering. *Measurement and Evaluation in Counseling and Development, 47*(2), 127-149.

Lumpe, A. T., & Beck, J. (1996). A profile of high school biology textbooks using scientific literacy recommendations. *The American Biology Teacher,* 147-153.

Martin, M.O. & Mullis, I.V.S. (Eds.). (2012). *Methods and procedures in TIMSS and PIRLS 2011.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist, 34*(4), 207-218.

Massachusetts Department of Elementary and Secondary Education. (2006). *Massachusetts Science and Technology/Engineering Curriculum Framework.* Retrieved from: http://www.doe.mass.edu/frameworks/scitech/1006.pdf

Massachusetts Department of Elementary and Secondary Education. (2016). *2016 Massachusetts Science and Technology/Engineering Curriculum Framework.* Retrieved from: http://www.doe.mass.edu/frameworks/scitech/2016-01.pdf

Massachusetts Department of Elementary and Secondary Education. (2015). *Massachusetts Comprehensive Assessment System: Science and Technology/Engineering (STE) test blueprints.* Retrieved from: http://www.doe.mass.edu/mcas/tdd/sci.html?section=testdesign

Massachusetts Department of Elementary and Secondary Education. (2016). *Percent of students at each achievement level for Newton.* Retrieved from: http://profiles.doe.mass.edu/mcas/achievement_level.aspx?linkid=32&orgcode=02070000&orgtypecode=5&

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Masters, G. N., & Wilson, M. (1997). *Developmental assessment.* Berkeley, CA: Berkeley Evaluation and Assessment Research Center.

McElhaney, K.W., deBarger, A.H., D'Angelo, C.M., Harris, C.J., Seeratan, K.L., & Stanford, T.M. (2015). Integrating Crosscutting Concepts into 3-Dimensional Scoring Rubrics. Paper presented at the *NARST Annual International Conference*, Chicago, IL.

Messick, S. (1988). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.). New York: Macmillan.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*(2), 133-161.

Mislevy, R. J. (2007). Validity by design. *Educational Researcher, 36*(8), 463-469.

Moskal, B. M. (2000). Scoring rubrics: What, when, and how? *Practical Assessment, Research, and Evaluation, 7*(3).

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386-422.

National Assessment Governing Board. (2007). *Science Assessment and Item Specifications for the 2009 National Assessment of Educational Progress.* Washington, D. C.: U. S. Department of Education.

National Assessment Governing Board. (2009). *Item Scaling Models.* Retrieved from https://nces.ed.gov/nationsreportcard/tdw/analysis/scaling_models.aspx

National Assessment Governing Board. (2010). *Science Framework for the 2011 National Assessment of Educational Progress.* Washington, D. C.: U. S. Department of Education.

National Assessment Governing Board. (2011). *Content of the Subject Area Scales.* Retrieved from https://nces.ed.gov/nationsreportcard/tdw/analysis/scaling_determination_naep.aspx

National Assessment Governing Board. (2016). Technical Notes on the Interactive Computer and Hands-On Tasks In Science. Retrieved from http://www.nationsreportcard.gov/science_2009/ict_tech_notes.asp

National Center for Education Statistics (2009). *NAEP Data Explorer* [Data file]. Retrieved from http://nces.ed.gov/nationsreportcard/naepdata/dataset.aspx

National Center for Education Statistics (2011). *The Nation's Report Card: Science 2009 (NCES 2011–451).* Washington, DC: Institute of Education Sciences, U.S. Department of Education.

National Center for Education Statistics. (2012). *The Nation's Report Card: Hands-on and interactive computer tasks from the 2009 science assessment.* Washington, DC: Institute of Education Sciences, U.S. Department of Education.

National Governors Association Center for Best Practices, Council of Chief State School Officers. (2010). *Common Core State Standards: Mathematics.* Washington, DC: National Governors Association Center for Best Practices, Council of Chief State School Officers.

National Research Council. (1996). *National science education standards.* Washington, DC: The National Academies Press.

National Research Council. (2011). *Incentives and test-based accountability in education*. Washington, DC: The National Academies Press.

National Research Council. (2000). *Inquiry and the national science education standards.* Washington, DC: The National Academies Press.

National Science Teachers' Association. (2004). *Position Statement: Scientific Inquiry.* Retrieved from http://www.nsta.org/docs/PositionStatement_ScientificInquiry.pdf

NGSS Lead States. (2013). *Next Generation Science Standards: For States, By States.* Washington, DC: The National Academies Press.

NGSS Lead States. (2014). *NGSS/CCSS-M Sample Classroom Assessment Tasks.* Retrieved from http://www.nextgenscience.org/classroom-sample-assessment-tasks

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231-259.

OECD. (2008). A profile of student performance in science. In *PISA 2006: Science Competencies for Tomorrow's World: Volume 1: Analysis* (p. 31-119). Paris, France: OECD.

OECD, (2012a). *PISA 2012 Technical Report.* Paris, France: OECD.

OECD. (2012b). *PISA 2015 Item Submission Guidelines: Scientific Literacy.* Paris, France: OECD.

OECD. (2013). *PISA 2015 Draft Science Framework.* Paris, France: OECD.

OECD. (2015). *PISA 2015 released field trial cognitive items.* Paris, France: OECD.

OECD. (2017) *PISA 2015 Technical Report*. Paris, France: OECD.

Osborne, J. F., Henderson, J. B., MacPherson, A., Szu, E., Wild, A., & Yao, S. (in press). The development and validation of a learning progression for argumentation in science. *Journal of Research in Science Teaching.*

Pea, R. D. (2004). The social and technological dimensions of scaffolding and related theoretical concepts for learning, education, and human activity. *The Journal of the Learning Sciences, 13*(3), 423-451.

Pellegrino, J. W. (2013). Proficiency in science: Assessment challenges and opportunities. *Science, 340*, 320-323.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: The National Academies Press.

Pellegrino, J. W., Wilson, M. R., Koenig, J. A., & Beatty, A. S. (2014). *Developing Assessments for the Next Generation Science Standards.* Washington, DC: The National Academies Press.

Penny, J. A., & Johnson, R. L. (2011). The accuracy of performance task scores after resolution of rater disagreement: A Monte Carlo study. *Assessing Writing, 16*(4), 221-236.

Peoples, S. (2012). *The Nature of Science Instrument – Elementary (NOSI-E): Using Rasch principles to develop a theoretically-grounded scale to measure elementary student understanding of the nature of science* (Unpublished doctoral dissertation). Boston College, Chestnut Hill, MA.

Quellmalz, E. S., Timms, M. J., Silberglitt, M. D., & Buckley, B. C. (2012). Science assessments for all: Integrating science simulations into balanced state science assessment systems. *Journal of Research in Science Teaching*,*49*(3), 363-393.

Quintana, C., Reiser, B. J., Davis, E. A., Krajcik, J., Fretz, E., Duncan, R.G., …Soloway, E., (2004). A scaffolding design framework for software to support science inquiry. *The Journal of the Learning Sciences, 13*(3), 337-386.

Rasch, G. (1960). *Probabilistic models for intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Rasch, G. (1980). *Probabilistic models for intelligence and attainment tests (expanded edition).* Chicago, IL: University of Chicago Press.

Reshetar, R. (2012). *AP Biology Exam* [PDF document]. Retrieved from https://research.collegeboard.org/sites/default/files/publications/

Resnick, D. P., & Resnick, L. B. (1996). Performance assessment and the multiple functions of educational measurement. In M. B. Kane & R. Mitchell (Eds.), *Implementing performance assessment: Promises, problems, and challenges* (23-38). Washington, DC: American Institutes for Research.

Roberts, L., Wilson, M., & Draney, K. (1997). *The SEPUP assessment system: An overview* (BEAR Report SA-97-1). Berkeley, CA: Berkeley Evaluation and Assessment Research Center.

Rossman, G. B., & Rallis, S. F. (2012). *Learning in the field: An introduction to qualitative research* (2nd ed.). Thousand Oaks, CA: Sage.

Russell, M. K., & Airasian, P. W. (2012). *Classroom assessment: Concepts and applications (7th edition).* New York, NY: McGraw-Hill.

Rutherford, F. J., & Ahlgren, A. (1989). *Science for all Americans: A Project 2061 report on literacy goals in science, mathematics, and technology.* Washington, DC: American Association for the Advancement of Science.

Ryan, K., Gannon-Slater, N., & Culbertson, M. J. (2012). Improving survey methods with cognitive interviews in small-and medium-scale evaluations. *American Journal of Evaluation, 33*(3), 414-430.

Scalise, K., & Gifford, B. (2006). Computer-based assessment in e-learning: A framework for constructing "intermediate constraint" questions and tasks for technology platforms. *The Journal of Technology, Learning, and Assessment, 4*(6).

Schwab, C. J. (2007). *What can we learn from PISA? Investigating PISA's approach to scientific literacy* (Unpublished doctoral dissertation). University of California, Berkeley, Berkeley, CA.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428.

Siegel, M. A., Nagle, B., & Barter, A. (2004) Development of an assessment instrument for middle school life science: Design considerations and consequences related to scoring student learning with rubrics. Paper presented at the *Annual Meeting of the American Educational Research Association*, San Diego, CA.

Smith, C. L. (2009). *Interview Protocol version 2-12.*

Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications of Research on Children's Learning for Standards and Assessment: A Proposed Learning Progression for Matter and the Atomic-Molecular Theory. *Measurement: Interdisciplinary Research & Perspective*, *4*(1-2), 1-98.

Songer, N. B., & Gotwals, A. W. (2012). Guiding explanation construction by children at the entry points of learning progressions. *Journal of Research in Science Teaching, 49*(2), 141-165.

Songer, N. B., Kelcey, B., & Gotwals, A. W. (2009). How and when does complex reasoning occur? Empirically driven development of a learning progression focused on complex reasoning about biodiversity. *Journal of Research in Science Teaching, 46*(6), 610-631.

Stanger-Hall, K. F. (2012). Multiple-choice exams: An obstacle for higher-level thinking in introductory science classes. *CBE – Life Sciences Education, 11*, 294-306.

Stenner, A. J., Smith, M., & Burdick, D. S. (1983). Toward a theory of construct definition. Journal of Educational Measurement, 20(4), 1-12.

Stone, C. A., Ye, F., Zhu, X., & Lane, S. (2010). Providing subscale scores for diagnostic information: A case study when the test is essentially unidimensional. *Applied Measurement in Education*, *23*, 63–86.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*, 589–617.

TERC. (2011). *The Inquiry Project: Seeing the world through a scientist's eyes.* Retrieved from: http://inquiryproject.terc.edu/.

TERC. (n.d.) *The inquiry curriculum and the standards for K-12 science education.* Retrieved from: http://www.doingnewsciencestandards.com/.

Verplanck, W. S. (1962). Unaware of where's awareness: Some verbal operants – notates, monents, and notants. *Journal of Personality, 30*(3), 130-158.

Vygotsky, L. S. (1978). *Mind in society: The development of higher metal processes* (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds.). Cambridge, MA: Harvard University Press.

Wang, C. (2015). On latent trait estimation in multidimensional compensatory item response models. *Psychometrika, 80*(2), 428-449.

Wang, W., & Chen, C. (2005). Item parameter recovery, standard error estimates, and fit statistis of the WINSTEPS program for the family of Rasch models. *Educational and Psychological Measurement, 65*(3), 376-404.

Wang, W., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement, 29*(2), 126-149.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427-450.

Weiss, I. R., Pasley, J. D., Smith, P. S., Banilower, E. R., & Heck, D. J. (2003). *Looking inside the classroom.* Chapel Hill, NC: Horizon Research, Inc.

WestEd & Council of Chief State School Officers. (2015a). *Science Assessment Item Collaborative Assessment Framework for the Next Generation Science Standards.* Washington, D.C.

WestEd & Council of Chief State School Officers. (2015b). *Science Assessment Item Collaborative High School Item Cluster Prototype for Assessment of the Next Generation Science Standards.* Washington, D.C.

WestEd & Council of Chief State School Officers. (2015c). *Science Assessment Item Collaborative Item Specifications Guidelines for the Next Generation Science Standards.* Washington, D.C.

Willis, G. B. (1999). Cognitive interviewing: A "How To" guide. Presented at the *Annual Meeting of the American Statistical Association.*

Wilson, M. (2005). *Constructing measures: An item response modeling approach.* New York, NY: Psychology Press.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching, 46*(6), 716-730.

Wilson, M. (2012). Responding to a challenge that learning progressions pose to measurement practice. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science: Current challenges and future directions.* Rotterdam, the Netherlands: Sense Publishing.

Wilson, M., & Draney, K. (2004). Some links between large-scale and classroom assessments: The case of the BEAR Assessment System. *Yearbook of the National Society for the Study of Education, 103*(2), 132-154.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13*(2), 181-208.

Wiser, M., Smith, C. L., & Doubler, S. (2012). Learning progressions as tools for curriculum development: Lessons from the Inquiry Project. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science: Current challenges and future directions.* Rotterdam, the Netherlands: Sense Publishing.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry, 17*(2), 89-100.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.

Wright, B. D., and Masters, G. N. (1982). *Rating scale analysis.* Chicago, IL: MESA Press

Zhang, J., & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika*, *64*, 213–249.

Appendix A.

Cognitive interview protocol.

Introduction:

I'm trying to write good test questions that students can understand, and I want your opinion to help me make these questions better. I am going to ask you to solve some test problems for me and then tell me about whether you understood the questions, what you thought about while you were solving the problems, and what was confusing for you. I'm not interested in whether you are right or wrong – I am just interested in what you think about the questions. Your answers won't be graded.

What you say is really important, so I am going to run this tape recorder to make sure I don't forget anything.

Do you have any questions?

First item:

Here is a test question. Read it quietly to yourself and write down your answer. Take as much time as you need and let me know when you are finished.

(Once the student indicates completion)

What did you think this question was asking you to do? (Repeat for multiple prompts, if applicable.)

Can you tell me what you were thinking about when you answered this question? (Repeat for multiple prompts, if applicable.)

Why did you pick that answer option?

Can you point to any words you didn't understand?

Can you point to any pictures that you didn't understand?

Was there anything else that was confusing about this question?

Other questions may be asked, based on specific observations:

Why did you choose not to respond to this question/part?

Why did you draw/write on this image?
Probes about their responses:

What do you mean by <insert word or phrase used in response>?

Other spontaneous probes may be utilized, depending on specific observations. Spontaneous probes may only ask about student understanding of the questions, interaction with item elements, their thought processes while answering, and their responses. Probes will not evaluate student comprehension of the tested material or anything else unrelated to the test item.

Repeat until 5 items are completed.

Conclusion:

Ok, we're finished. Nice job. Thank you so much for helping me out today!

Appendix B

Frequency of all Sources of Student Confusion During Cognitive Interviews

Table B.1

Round 1: Frequency of All Sources of Student Confusion

	Multiple-prompt explicit multidimensional		Single-prompt explicit multidimensional		Single-prompt implicit multidimensional	
Issue	Count	Percentage	Count	Percentage	Count	Percentage
Misunderstanding or misinterpretation of critical piece of task	10	23%	11	26%	11	25%
Overlooks a critical piece of information	1	2%	5	12%	1	2%
Unfamiliarity with task context	1	2%	2	5%	1	2%
Misunderstanding of visuals	8	19%	7	16%	3	7%
Unfamiliarity with item vocabulary (non-scientific)	3	7%	3	7%	2	5%
Unfamiliarity with measurement unit	7	16%	7	16%	6	14%
Unfamiliarity with measurement tool	5	12%	2	5%	7	16%
Confused by item layout	1	2%	2	5%	3	7%
Misunderstanding of key concepts	12	28%	4	9%	4	9%
Unfamiliarity with measurement calculation	6	14%	3	7%	5	11%
Provides "correct" answer by avoiding intended task	0	0%	2	5%	0	0%

Table B.2

Round 2: Frequency of All Sources of Student Confusion

Issue	Constructed response/ Multiple prompt		Selected response/ Multiple prompt		Selected response/ Single prompt	
	Count	Percentage	Count	Percentage	Count	Percentage
Misunderstanding or misinterpretation of critical piece of task	8	17.02%	21	30.00%	15	25.42%
Unfamiliar and/or misunderstanding of key concept	9	19.15%	20	28.57%	3	5.08%
Unfamiliarity with item vocabulary (non-scientific)	2	4.26%	0	0.00%	0	0.00%
Unfamiliarity with measurement calculation	8	17.02%	7	10.00%	4	6.78%
Unfamiliarity with measurement unit	2	4.26%	3	4.29%	7	11.86%
Misunderstanding of visuals	5	10.64%	7	10.00%	3	5.08%
Answer options don't reflect student understanding	0	0.00%	8	11.43%	13	22.03%
Overwhelmed by amount of information and/or answer choices	1	2.13%	6	8.57%	4	6.78%
Alternative explanation based on extraneous factor	4	8.51%	4	5.71%	4	6.78%
Provide "correct" answer by avoiding intended task	9	19.15%	11	15.71%	10	16.95%
Expresses preference for selected response argument	0	0.00%	13	18.57%	2	3.39%
Expresses preference for written argument	1	2.13%	5	7.14%	1	1.69%
Answer options lead student to rethink answer	0	0.00%	9	12.86%	9	15.25%
Confused by item layout	2	4.26%	0	0.00%	0	0.00%

	Constructed response/ Multiple prompt		Selected response/ Multiple prompt		Selected response/ Single prompt	
Issue	Count	Percentage	Count	Percentage	Count	Percentage
Large/small scale makes problem difficult to think about	1	2.13%	0	0.00%	4	6.78%
Misunderstanding of answer options	0	0.00%	2	2.86%	1	1.69%
Overlooks a critical piece of information	5	10.64%	3	4.29%	5	8.47%
Unfamiliarity with task context	3	6.38%	3	4.29%	2	3.39%

Appendix C

Interrater Reliability for All Items Under Both Rubrics, Measured by Intraclass Correlation Coefficients (Consistency Measure)

Table C.1

Interrater Reliability Based on the Multidimensional Rubric

Item	ICC - SPQ	ICC - Matter	ICC - Argument
1A - Ana's block of clay	0.999	0.965	0.601
1B - Ana's block of clay	0.933	0.999	0.654
2A - Beth's rock – 1	0.981	0.999	0.5
2A - Beth's rock – 2		0.969	
2B - Beth's rock – 1	0.957	0.999	0.744
2B - Beth's rock – 2		0.999	
3A - Romita's cylinders	0.999	0.999	0.511
3B - Romita's cylinders		0.936	0.306
4A - Kevin's coffee	0.999	0.972	0.694
4B - Kevin's coffee	0.999	0.945	0.668
5A - Brass and aluminum – 1	0.999	0.999	0.637
5A - Brass and aluminum – 2	0.999		
5B - Brass and aluminum – 1	0.956	0.896	0.586
5B - Brass and aluminum – 2	0.917		
6A - Carol's butter	0.999	0.999	0.458
6B - Carol's butter	0.86	0.999	0.215
7A - Box of blocks – 1	0.999	0.999	0.411
7A - Box of blocks – 2	0.999		
7A - Box of blocks – 3	0.963		
7B - Box of blocks	0.946	0.742	0.593
8A - Grain of sugar – 1	0.929	0.999	0.581
8A - Grain of sugar – 2	0.732	0.999	0.469
8B - Grain of sugar – 1	0.878	0.903	0.655
8B - Grain of sugar – 2	0.822	0.999	0.58
9A - Nate's can of soda	0.971	0.999	0.669
9B - Nate's can of soda	0.379	0.999	0.844
10 - Drops of water – 1	0.999	0.999	0.775
10 - Drops of water – 2	0.969	0.999	0.786
11 - Ken's cubes	0.942	0.999	0.673

Table C.2

Interrater Reliability Based on the Holistic Rubric

Item	ICC
1A - Ana's block of clay	0.804
1B - Ana's block of clay	0.91
2A - Beth's rock	0.753
2B - Beth's rock	0.657
3A - Romita's cylinders	0.746
3B - Romita's cylinders	0.771
4A - Kevin's coffee	0.734
4B - Kevin's coffee	0.714
5A - Brass and aluminum	0.749
5B - Brass and aluminum	0.757
6A - Carol's butter	0.741
6B - Carol's butter	0.798
7A - Box of blocks	0.735
7B - Box of blocks	0.873
8A - Grain of sugar	0.526
8B - Grain of sugar	0.735
9A - Nate's can of soda	0.648
9B - Nate's can of soda	0.724
10 - Drops of water	0.749
11 - Ken's cubes	0.851

Appendix D

Item Difficulty Estimates and Fit Statistics for All Items and Models

Table D.1

Parameter Estimates and Fit: Unidimensional Model, Multidimensional Rubric

Parameter	Parameter estimate	Standard error	Unweighted Mean Square	Confidence Interval	T	Weighted Mean Square	Confidence Interval	T
Argument item1	-0.87	0.23	1.42	(0.73, 1.27)	2.8	1.57	(0.67, 1.33)	2.9
Argument item2	-0.69	0.16	1.19	(0.71, 1.29)	1.3	1.34	(0.68, 1.32)	1.9
Argument item3	-0.85	0.24	1.14	(0.75, 1.25)	1.0	1.13	(0.74, 1.26)	1.0
Argument item4	-1.19	0.23	1.60	(0.74, 1.26)	3.9	1.75	(0.71, 1.29)	4.2
Argument item5	-0.85	0.19	1.18	(0.76, 1.24)	1.5	1.29	(0.73, 1.27)	2.0
Argument item6	-0.85	0.18	1.23	(0.75, 1.25)	1.7	1.17	(0.75, 1.25)	1.3
Argument item7	-1.10	0.20	1.11	(0.75, 1.25)	0.9	1.30	(0.71, 1.29)	1.9
Argument item8	-1.78	0.24	1.36	(0.77, 1.23)	2.8	1.47	(0.73, 1.27)	3.0
Argument item9	-1.20	0.27	1.36	(0.74, 1.26)	2.5	1.42	(0.71, 1.29)	2.5
Argument item10	-1.24	0.20	1.39	(0.76, 1.24)	2.9	1.56	(0.72, 1.28)	3.4
Argument item11	-1.16	0.25	1.73	(0.74, 1.26)	4.6	1.25	(0.66, 1.34)	1.4
Argument item12	-0.86	0.25	1.00	(0.75, 1.25)	0.0	1.04	(0.71, 1.29)	0.3
Argument item13	-0.98	0.20	1.12	(0.75, 1.25)	0.9	1.25	(0.69, 1.31)	1.5
Argument item14	-0.93	0.16	1.48	(0.76, 1.24)	3.4	1.44	(0.75, 1.25)	3.0
Argument item15A	-1.18	0.21	1.10	(0.75, 1.25)	0.8	1.33	(0.69, 1.31)	1.9
Argument item15B	-0.87	0.27	1.14	(0.74, 1.26)	1.1	1.37	(0.70, 1.30)	2.2
Argument item16A	-0.84	0.14	1.33	(0.74, 1.26)	2.3	1.29	(0.73, 1.27)	2.0
Argument item16B	-0.29	0.13	1.50	(0.73, 1.27)	3.2	1.48	(0.73, 1.27)	3.1

Parameter	Parameter estimate	Standard error	Unweighted Mean Square	Confidence Interval	T	Weighted Mean Square	Confidence Interval	T
Argument item17	-1.23	0.27	1.05	(0.76, 1.24)	0.4	1.39	(0.71, 1.29)	2.4
Argument item18	-1.79	0.27	1.47	(0.74, 1.26)	3.2	1.53	(0.70, 1.30)	3.0
Argument item19	-0.74	0.14	1.18	(0.84, 1.16)	2.1	1.37	(0.81, 1.19)	3.4
Argument item20	-1.03	0.16	1.00	(0.84, 1.16)	0.0	1.32	(0.79, 1.21)	2.7
Argument rater1	0.00	0.16	1.08	(0.85, 1.15)	1.0	1.07	(0.83, 1.17)	0.8
Argument rater2	-0.37	0.13	1.05	(0.85, 1.15)	0.6	1.07	(0.83, 1.17)	0.8
Argument rater3	-0.36	0.15	1.15	(0.85, 1.15)	2.0	1.21	(0.83, 1.17)	2.3
Argument rater4	0.21	0.09	1.28	(0.85, 1.15)	3.5	1.37	(0.84, 1.16)	4.1
Argument rater5	0.29	0.15	1.11	(0.85, 1.15)	1.4	1.04	(0.84, 1.16)	0.5
Argument rater6	0.17	0.11	0.94	(0.85, 1.15)	-0.8	0.95	(0.83, 1.17)	-0.6
Argument item1 x step0to1	0.26	0.57	1.34	(0.73, 1.27)	2.3	1.49	(0.78, 1.22)	3.9
Argument item1 x step1to2	-2.03	0.48	1.32	(0.73, 1.27)	2.2	1.45	(0.79, 1.21)	3.8
Argument item2 x step0to1	0.10	0.34	1.24	(0.71, 1.29)	1.5	1.30	(0.77, 1.23)	2.4
Argument item2 x step1to2	-1.48	0.32	1.39	(0.71, 1.29)	2.4	1.44	(0.79, 1.21)	3.6
Argument item3 x step0to1	-1.85	0.45	1.14	(0.75, 1.25)	1.1	1.21	(0.78, 1.22)	1.8
Argument item3 x step1to2	0.15	0.29	1.11	(0.75, 1.25)	0.9	1.16	(0.82, 1.18)	1.7
Argument item4 x step0to1	-0.68	0.46	1.53	(0.74, 1.26)	3.5	1.65	(0.80, 1.20)	5.4
Argument item4 x step1to2	-0.76	0.34	1.47	(0.74, 1.26)	3.2	1.57	(0.81, 1.19)	5.1
Argument item5 x step0to1	-0.35	0.39	0.92	(0.76, 1.24)	-0.6	0.99	(0.81, 1.19)	-0.1
Argument item5 x step1to2	-0.38	0.33	0.79	(0.76, 1.24)	-1.8	0.95	(0.79, 1.21)	-0.5
Argument item6 x step0to1	-0.93	0.30	1.23	(0.75, 1.25)	1.8	1.28	(0.80, 1.20)	2.5
Argument item6 x step1to2	-0.31	0.25	1.14	(0.75, 1.25)	1.1	1.19	(0.82, 1.18)	2.0
Argument item7 x step0to1	-0.37	0.45	1.23	(0.75, 1.25)	1.7	1.33	(0.80, 1.20)	2.9
Argument item7 x step1to2	-1.40	0.38	1.18	(0.75, 1.25)	1.4	1.26	(0.82, 1.18)	2.6
Argument item8 x step0to1	-0.91	0.48	1.43	(0.77, 1.23)	3.2	1.53	(0.82, 1.18)	5.0

Parameter	Parameter estimate	Standard error	Unweighted Mean Square	Confidence Interval	T	Weighted Mean Square	Confidence Interval	T
Argument item8 x step1to2	-0.67	0.34	1.39	(0.77, 1.23)	3.0	1.50	(0.83, 1.17)	5.0
Argument item9 x step0to1	-1.51	0.54	1.11	(0.74, 1.26)	0.9	1.15	(0.78, 1.22)	1.3
Argument item9 x step1to2	-0.82	0.36	1.10	(0.74, 1.26)	0.8	1.14	(0.80, 1.20)	1.3
Argument item10 x step0to1	-0.27	0.39	1.15	(0.76, 1.24)	1.2	1.20	(0.82, 1.18)	2.0
Argument item10 x step1to2	-1.19	0.32	1.20	(0.76, 1.24)	1.6	1.28	(0.83, 1.17)	3.0
Argument item11 x step0to1	-0.77	0.50	1.01	(0.74, 1.26)	0.1	1.26	(0.77, 1.23)	2.0
Argument item11 x step1to2	-0.39	0.37	1.00	(0.74, 1.26)	0.0	1.34	(0.76, 1.24)	2.6
Argument item12 x step0to1	-0.59	0.64	0.71	(0.75, 1.25)	-2.6	0.96	(0.72, 1.28)	-0.3
Argument item12 x step1to2	-0.93	0.56	0.93	(0.75, 1.25)	-0.5	0.94	(0.82, 1.18)	-0.7
Argument item13 x step0to1	-0.82	0.44	0.89	(0.75, 1.25)	-0.8	1.17	(0.73, 1.27)	1.2
Argument item13 x step1to2	-1.43	0.36	1.12	(0.75, 1.25)	0.9	1.20	(0.80, 1.20)	1.9
Argument item14 x step0to1	0.12	0.36	1.14	(0.76, 1.24)	1.1	1.20	(0.82, 1.18)	2.1
Argument item14 x step1to2	-1.18	0.32	1.11	(0.76, 1.24)	0.9	1.22	(0.81, 1.19)	2.2
Argument item15A x step0to1	-0.96	0.41	1.21	(0.75, 1.25)	1.6	1.44	(0.77, 1.23)	3.3
Argument item15A x step1to2	-0.32	0.33	1.15	(0.75, 1.25)	1.2	1.23	(0.79, 1.21)	2.0
Argument item15B x step0to1	-2.36	0.51	1.03	(0.74, 1.26)	0.2	1.34	(0.66, 1.34)	1.9
Argument item15B x step1to2	-0.48	0.32	1.07	(0.74, 1.26)	0.6	1.17	(0.77, 1.23)	1.5
Argument item16A x step0to1	-0.05	0.31	1.25	(0.74, 1.26)	1.8	1.32	(0.81, 1.19)	2.9
Argument item16A x step1to2	-0.36	0.29	1.18	(0.74, 1.26)	1.3	1.24	(0.79, 1.21)	2.1
Argument item16B x step0to1	-0.05	0.26	1.18	(0.73, 1.27)	1.3	1.19	(0.80, 1.20)	1.8
Argument item16B x step1to2	-0.67	0.26	1.02	(0.73, 1.27)	0.2	1.03	(0.78, 1.22)	0.3
Argument item17 x step0to1	-1.02	0.56	1.05	(0.76, 1.24)	0.4	1.29	(0.78, 1.22)	2.4
Argument item17 x step1to2	-0.95	0.39	1.05	(0.76, 1.24)	0.5	1.25	(0.79, 1.21)	2.3
Argument item18 x step0to1	-0.13	0.66	1.56	(0.74, 1.26)	3.7	1.69	(0.80, 1.20)	5.7
Argument item18 x step1to2	-1.71	0.52	1.50	(0.74, 1.26)	3.4	1.62	(0.81, 1.19)	5.3

Parameter	Parameter estimate	Standard error	Unweighted Mean Square	Confidence Interval	T	Weighted Mean Square	Confidence Interval	T
Argument item19 x step0to1	-0.75	0.27	1.18	(0.84, 1.16)	2.1	1.31	(0.86, 1.14)	3.8
Argument item19 x step1to2	-0.44	0.24	1.16	(0.84, 1.16)	1.9	1.24	(0.89, 1.11)	4.0
Argument item20 x step0to1	-0.06	0.34	0.99	(0.84, 1.16)	0.0	1.08	(0.89, 1.11)	1.4
Argument item20 x step1to2	-1.39	0.29	0.95	(0.84, 1.16)	-0.6	1.02	(0.89, 1.11)	0.4
Argument item1 x rater1	0.15	0.17	0.67	(0.85, 1.15)	-5.0	0.67	(0.82, 1.18)	-4.0
Argument item3 x rater2	0.00	0.18	0.69	(0.85, 1.15)	-4.6	0.65	(0.82, 1.18)	-4.2
Argument item7 x rater1	0.09	0.15	0.69	(0.85, 1.15)	-4.5	0.62	(0.79, 1.21)	-4.1
Argument item8 x rater4	0.04	0.17	0.66	(0.83, 1.17)	-4.3	0.64	(0.80, 1.20)	-4.0
Argument item9 x rater2	-0.04	0.22	0.62	(0.80, 1.20)	-4.3	0.60	(0.78, 1.22)	-4.2
Argument item10 x rater6	0.14	0.13	0.84	(0.80, 1.20)	-1.7	0.84	(0.78, 1.22)	-1.5
Argument item11 x rater2	0.13	0.19	0.73	(0.85, 1.15)	-4.0	0.72	(0.80, 1.20)	-3.0
Argument item12 x rater2	-0.34	0.16	0.79	(0.84, 1.16)	-2.8	0.73	(0.82, 1.18)	-3.3
Argument item13 x rater1	0.05	0.15	0.81	(0.85, 1.15)	-2.6	0.77	(0.82, 1.18)	-2.7
Argument item14 x rater4	-0.17	0.12	0.73	(0.80, 1.20)	-2.8	0.64	(0.77, 1.23)	-3.5
Argument item15A x rater3	0.17	0.16	0.68	(0.85, 1.15)	-4.5	0.68	(0.79, 1.21)	-3.4
Argument item15B x rater3	0.10	0.21	0.66	(0.85, 1.15)	-4.8	0.66	(0.80, 1.20)	-3.7
Argument item16A x rater4	0.32	0.11	0.36	(0.74, 1.26)	-6.5	0.30	(0.73, 1.27)	-7.1
Argument item17 x rater1	-0.29	0.20	0.66	(0.85, 1.15)	-5.1	0.65	(0.81, 1.19)	-4.2
Argument item18 x rater4	-0.17	0.18	0.83	(0.84, 1.16)	-2.2	0.72	(0.82, 1.18)	-3.3
Argument item19 x rater2	0.16	0.12	0.71	(0.84, 1.16)	-4.0	0.64	(0.80, 1.20)	-4.1
Argument item1 x step0to1 x rater1	0.15	0.57	0.66	(0.73, 1.27)	-2.9	0.65	(0.78, 1.22)	-3.6
Argument item1 x step1to2 x rater1	-0.26	0.48	0.70	(0.73, 1.27)	-2.4	0.69	(0.79, 1.21)	-3.2
Argument item2 x step0to1 x rater5	-0.09	0.34	0.69	(0.71, 1.29)	-2.3	0.69	(0.77, 1.23)	-3.0
Argument item2 x step1to2 x rater5	0.20	0.32	0.60	(0.71, 1.29)	-3.1	0.56	(0.79, 1.21)	-4.8
Argument item3 x step0to1 x rater2	0.00	0.45	0.73	(0.75, 1.25)	-2.3	0.66	(0.78, 1.22)	-3.5

Parameter	Parameter estimate	Standard error	Unweighted Mean Square	Confidence Interval	T	Weighted Mean Square	Confidence Interval	T
Argument item3 x step1to2 x rater2	-0.08	0.29	0.81	(0.75, 1.25)	-1.6	0.78	(0.82, 1.18)	-2.5
Argument item4 x step0to1 x rater2	-0.62	0.46	0.56	(0.74, 1.26)	-4.0	0.41	(0.80, 1.20)	-7.5
Argument item4 x step1to2 x rater2	0.22	0.34	0.55	(0.74, 1.26)	-4.1	0.44	(0.81, 1.19)	-7.4
Argument item5 x step0to1 x rater1	-0.35	0.39	0.87	(0.76, 1.24)	-1.0	0.91	(0.81, 1.19)	-0.9
Argument item5 x step1to2 x rater1	0.62	0.33	0.82	(0.76, 1.24)	-1.5	0.98	(0.79, 1.21)	-0.2
Argument item6 x step0to1 x rater6	0.77	0.30	0.87	(0.75, 1.25)	-1.1	0.80	(0.80, 1.20)	-2.1
Argument item6 x step1to2 x rater6	-0.69	0.25	0.83	(0.75, 1.25)	-1.4	0.80	(0.82, 1.18)	-2.3
Argument item7 x step0to1 x rater1	-0.35	0.45	0.67	(0.75, 1.25)	-2.9	0.64	(0.80, 1.20)	-4.0
Argument item7 x step1to2 x rater1	0.39	0.38	0.72	(0.75, 1.25)	-2.4	0.69	(0.82, 1.18)	-3.8
Argument item8 x step0to1 x rater4	-0.23	0.48	0.68	(0.77, 1.23)	-3.0	0.53	(0.82, 1.18)	-6.2
Argument item8 x step1to2 x rater4	0.12	0.34	0.65	(0.77, 1.23)	-3.3	0.51	(0.83, 1.17)	-7.0
Argument item9 x step0to1 x rater2	0.62	0.54	0.77	(0.74, 1.26)	-1.8	0.76	(0.78, 1.22)	-2.3
Argument item9 x step1to2 x rater2	-0.20	0.36	0.83	(0.74, 1.26)	-1.3	0.81	(0.80, 1.20)	-2.0
Argument item10 x step0to1 x rater6	0.45	0.39	0.81	(0.76, 1.24)	-1.7	0.75	(0.82, 1.18)	-2.9
Argument item10 x step1to2 x rater6	-0.48	0.32	0.72	(0.76, 1.24)	-2.5	0.67	(0.83, 1.17)	-4.2
Argument item11 x step0to1 x rater2	-0.59	0.50	0.54	(0.74, 1.26)	-4.1	0.61	(0.77, 1.23)	-3.8
Argument item11 x step1to2 x rater2	-0.17	0.37	0.46	(0.74, 1.26)	-5.1	0.57	(0.76, 1.24)	-4.2
Argument item12 x step0to1 x rater2	1.45	0.64	0.72	(0.75, 1.25)	-2.5	0.97	(0.72, 1.28)	-0.2
Argument item12 x step1to2 x rater2	-0.57	0.56	1.00	(0.75, 1.25)	0.1	1.02	(0.82, 1.18)	0.2
Argument item13 x step0to1 x rater1	0.76	0.44	0.71	(0.75, 1.25)	-2.5	0.86	(0.73, 1.27)	-1.0
Argument item13 x step1to2 x rater1	-0.43	0.36	0.85	(0.75, 1.25)	-1.2	0.84	(0.80, 1.20)	-1.6
Argument item14 x step0to1 x rater4	-0.07	0.35	0.82	(0.76, 1.24)	-1.5	0.79	(0.82, 1.18)	-2.3
Argument item14 x step1to2 x rater4	-0.55	0.32	0.77	(0.76, 1.24)	-2.0	0.80	(0.81, 1.19)	-2.2
Argument item15A x step0to1 x rater3	0.42	0.41	0.61	(0.75, 1.25)	-3.5	0.63	(0.77, 1.23)	-3.5
Argument item15A x step1to2 x rater3	0.65	0.33	0.81	(0.75, 1.25)	-1.6	0.80	(0.79, 1.21)	-2.0

Parameter	Parameter estimate	Standard error	Unweighted Mean Square	Confidence Interval	T	Weighted Mean Square	Confidence Interval	T
Argument item15B x step0to1 x rater3	0.40	0.51	0.57	(0.74, 1.26)	-3.8	0.78	(0.66, 1.34)	-1.3
Argument item15B x step1to2 x rater3	0.44	0.32	0.79	(0.74, 1.26)	-1.6	0.81	(0.77, 1.23)	-1.7
Argument item16A x step0to1 x rater4	0.83	0.31	0.68	(0.74, 1.26)	-2.7	0.60	(0.81, 1.19)	-4.7
Argument item16A x step1to2 x rater4	-1.17	0.29	0.80	(0.74, 1.26)	-1.6	0.78	(0.79, 1.21)	-2.2
Argument item16B x step0to1 x rater4	0.47	0.26	0.90	(0.73, 1.27)	-0.7	0.85	(0.80, 1.20)	-1.5
Argument item16B x step1to2 x rater4	-0.70	0.26	0.99	(0.73, 1.27)	0.0	1.03	(0.78, 1.22)	0.3
Argument item17 x step0to1 x rater1	0.02	0.56	0.62	(0.76, 1.24)	-3.6	0.67	(0.78, 1.22)	-3.3
Argument item17 x step1to2 x rater1	0.66	0.39	0.64	(0.76, 1.24)	-3.3	0.70	(0.79, 1.21)	-3.2
Argument item18 x step0to1 x rater4	-0.02	0.66	0.53	(0.74, 1.26)	-4.3	0.39	(0.80, 1.20)	-8.0
Argument item18 x step1to2 x rater4	0.20	0.52	0.56	(0.74, 1.26)	-4.0	0.43	(0.81, 1.19)	-7.4
Argument item19 x step0to1 x rater2	0.25	0.27	0.87	(0.84, 1.16)	-1.7	0.77	(0.86, 1.14)	-3.3
Argument item19 x step1to2 x rater2	0.53	0.24	0.87	(0.84, 1.16)	-1.6	0.81	(0.89, 1.11)	-3.7
Argument item20 x step0to1 x rater3	0.71	0.33	0.83	(0.84, 1.16)	-2.2	0.83	(0.89, 1.11)	-3.2
Argument item20 x step1to2 x rater3	-0.41	0.29	0.88	(0.84, 1.16)	-1.6	0.89	(0.89, 1.11)	-2.0
SPQ item1	-1.07	0.21	1.05	(0.76, 1.24)	0.4	1.01	(0.81, 1.19)	0.1
SPQ item2	-0.45	0.23	1.01	(0.72, 1.28)	0.1	0.98	(0.84, 1.16)	-0.3
SPQ item3	-0.46	0.13	1.08	(0.76, 1.24)	0.7	1.07	(0.81, 1.19)	0.8
SPQ item4	-0.25	0.13	1.06	(0.74, 1.26)	0.5	1.07	(0.80, 1.20)	0.7
SPQ item5	-0.71	0.20	1.00	(0.76, 1.24)	0.0	1.01	(0.85, 1.15)	0.1
SPQ item6	-1.38	0.28	1.16	(0.70, 1.30)	1.0	1.03	(0.70, 1.30)	0.3
SPQ item7	-1.41	0.23	0.93	(0.76, 1.24)	-0.5	0.99	(0.75, 1.25)	0.0
SPQ item8	-1.45	0.22	0.80	(0.77, 1.23)	-1.8	0.91	(0.78, 1.22)	-0.8
SPQ item9	-2.26	0.29	0.98	(0.75, 1.25)	-0.1	0.99	(0.59, 1.41)	0.0
SPQ item10	-1.77	0.24	1.02	(0.77, 1.23)	0.2	1.00	(0.71, 1.29)	0.0
SPQ item11	-1.78	0.25	0.82	(0.75, 1.25)	-1.5	0.93	(0.70, 1.30)	-0.4

Parameter	Parameter estimate	Standard error	Unweighted Mean Square	Confidence Interval	T	Weighted Mean Square	Confidence Interval	T
SPQ item12	-0.40	0.23	0.91	(0.71, 1.29)	-0.5	0.95	(0.86, 1.14)	-0.6
SPQ item13	0.64	0.17	0.93	(0.76, 1.24)	-0.6	0.92	(0.81, 1.19)	-0.8
SPQ item14	2.59	0.31	0.70	(0.76, 1.25)	-2.7	0.94	(0.54, 1.46)	-0.2
SPQ item15A	0.43	0.16	0.90	(0.68, 1.32)	-0.6	0.95	(0.76, 1.24)	-0.4
SPQ item15B	1.35	0.19	0.74	(0.70, 1.30)	-1.8	0.86	(0.60, 1.40)	-0.6
SPQ item16A	-0.16	0.15	0.75	(0.71, 1.29)	-1.8	0.78	(0.78, 1.22)	-2.0
SPQ item16B	0.37	0.14	0.81	(0.70, 1.30)	-1.2	0.88	(0.79, 1.21)	-1.2
SPQ item17	-0.16	0.19	0.90	(0.76, 1.24)	-0.8	0.92	(0.89, 1.11)	-1.6
SPQ item18	-0.42	0.21	1.09	(0.74, 1.26)	0.7	1.06	(0.87, 1.13)	0.9
SPQ item19	-1.13	0.12	0.89	(0.84, 1.16)	-1.4	0.92	(0.85, 1.15)	-1.1
SPQ item20	-1.02	0.14	0.89	(0.84, 1.16)	-1.4	0.94	(0.88, 1.12)	-1.1
SPQ item3 x step 0to1	0.41	0.22	0.87	(0.76, 1.24)	-1.0	0.96	(0.75, 1.25)	-0.3
SPQ item4 x step0to1	0.61	0.25	0.99	(0.74, 1.26)	0.0	1.00	(0.69, 1.31)	0.0
SPQ item13 x step0to1	-0.61	0.28	0.92	(0.76, 1.24)	-0.6	0.95	(0.81, 1.19)	-0.5
SPQ item13 x step1to2	-1.77	0.29	1.00	(0.76, 1.24)	0.0	0.99	(0.88, 1.12)	-0.1
SPQ item15A x step0to1	0.77	0.32	1.15	(0.68, 1.32)	1.0	1.05	(0.58, 1.42)	0.3
SPQ item15B x step0to1	1.12	0.42	0.85	(0.70, 1.30)	-1.0	0.98	(0.37, 1.63)	0.0
SPQ item16A x step0to1	-0.03	0.24	0.97	(0.71, 1.29)	-0.1	0.98	(0.81, 1.19)	-0.1
SPQ item16B x step0to1	0.91	0.31	0.96	(0.70, 1.30)	-0.2	0.99	(0.57, 1.43)	0.0
SPQ item19 x step0to1	-0.74	0.15	0.89	(0.84, 1.16)	-1.4	0.92	(0.93, 1.07)	-2.2
Matter item1	-0.43	0.13	1.06	(0.74, 1.26)	0.5	1.07	(0.80, 1.20)	0.7
Matter item2	-0.97	0.20	0.98	(0.69, 1.31)	-0.1	1.07	(0.56, 1.44)	0.4
Matter item3	0.32	0.13	0.99	(0.74, 1.26)	-0.1	1.01	(0.80, 1.20)	0.1
Matter item4	0.31	0.13	1.06	(0.73, 1.27)	0.5	1.08	(0.81, 1.19)	0.8
Matter item5	-0.22	0.12	0.88	(0.76, 1.24)	-1.0	0.96	(0.82, 1.18)	-0.4

Parameter	Parameter estimate	Standard error	Unweighted Mean Square	Confidence Interval	T	Weighted Mean Square	Confidence Interval	T
Matter item6	-0.22	0.12	0.85	(0.75, 1.25)	-1.2	0.90	(0.83, 1.17)	-1.2
Matter item7	0.57	0.12	1.39	(0.75, 1.25)	2.8	1.30	(0.80, 1.20)	2.8
Matter item8	0.20	0.11	1.15	(0.77, 1.23)	1.2	1.14	(0.83, 1.17)	1.5
Matter item9	-0.22	0.12	1.19	(0.75, 1.25)	1.5	1.15	(0.82, 1.18)	1.6
Matter item10	-0.01	0.11	1.17	(0.76, 1.24)	1.4	1.18	(0.84, 1.16)	2.1
Matter item11	0.34	0.12	1.26	(0.75, 1.25)	1.9	1.17	(0.80, 1.20)	1.6
Matter item12	0.01	0.12	1.08	(0.74, 1.26)	0.6	1.07	(0.84, 1.16)	0.8
Matter item13	1.61	0.25	0.86	(0.76, 1.24)	-1.1	0.91	(0.75, 1.25)	-0.7
Matter item14	0.75	0.16	0.98	(0.76, 1.24)	-0.1	0.97	(0.80, 1.20)	-0.2
Matter item15	-0.03	0.12	1.09	(0.74, 1.26)	0.7	1.03	(0.82, 1.18)	0.3
Matter item16	-0.95	0.19	0.85	(0.69, 1.31)	-1.0	1.01	(0.59, 1.41)	0.1
Matter item17	0.25	0.26	0.99	(0.76, 1.24)	-0.1	0.97	(0.83, 1.17)	-0.3
Matter item18	-0.84	0.14	0.82	(0.74, 1.26)	-1.4	0.94	(0.74, 1.26)	-0.4
Matter item19	-0.20	0.08	0.96	(0.85, 1.15)	-0.5	0.98	(0.89, 1.11)	-0.4
Matter item20	-0.45	0.08	1.12	(0.84, 1.16)	1.5	1.06	(0.88, 1.12)	0.9
Matter item1 step0to1	1.65	0.37	1.17	(0.74, 1.26)	1.3	1.02	(0.38, 1.62)	0.2
Matter item2 step0to1	2.37	0.73	1.16	(0.69, 1.31)	1.0	1.00	(0.00, 2.32)	0.2
Matter item3 step0to1	0.78	0.26	1.03	(0.74, 1.26)	0.3	1.01	(0.65, 1.35)	0.1
Matter item4 step0to1	1.20	0.31	1.04	(0.73, 1.27)	0.3	1.01	(0.53, 1.47)	0.1
Matter item5 step0to1	1.04	0.26	0.94	(0.76, 1.24)	-0.4	0.99	(0.62, 1.38)	0.0
Matter item6 step0to1	1.28	0.29	0.90	(0.75, 1.25)	-0.7	0.98	(0.55, 1.45)	0.0
Matter item7 step0to1	2.50	0.51	1.06	(0.75, 1.25)	0.5	1.00	(0.07, 1.93)	0.2
Matter item8 step0to1	2.41	0.46	1.23	(0.77, 1.23)	1.9	1.01	(0.17, 1.83)	0.2
Matter item9 step0to1	1.24	0.29	1.00	(0.75, 1.25)	0.0	1.00	(0.55, 1.45)	0.1
Matter item10 step0to1	4.04	1.00	0.91	(0.76, 1.24)	-0.7	1.00	(0.00, 2.94)	0.3

Parameter	Parameter estimate	Standard error	Unweighted Mean Square	Confidence Interval	T	Weighted Mean Square	Confidence Interval	T
Matter item11 step0to1	2.73	0.59	0.87	(0.75, 1.25)	-1.0	1.00	(0.00, 2.09)	0.2
Matter item 12 step0to1	3.89	1.00	0.89	(0.74, 1.26)	-0.9	1.00	(0.00, 2.93)	0.3
Matter item13 step0to1	-0.94	0.28	0.95	(0.76, 1.24)	-0.4	0.98	(0.85, 1.15)	-0.2
Matter item14 step0to1	-0.89	0.19	0.96	(0.76, 1.24)	-0.3	0.97	(0.94, 1.06)	-1.1
Matter item15 step0to1	2.23	0.46	1.18	(0.74, 1.26)	1.4	1.01	(0.18, 1.82)	0.2
Matter item16 step0to1	1.70	0.53	0.65	(0.69, 1.31)	-2.5	0.96	(0.11, 1.89)	0.1
Matter item17 step0to1	-1.07	0.44	0.98	(0.76, 1.24)	-0.1	0.98	(0.86, 1.14)	-0.3
Matter item18 step0to1	2.55	0.59	0.84	(0.74, 1.26)	-1.2	1.00	(0.00, 2.08)	0.2
Matter item19 step0to1	1.26	0.18	1.03	(0.85, 1.15)	0.4	1.00	(0.72, 1.28)	0.1
Matter item20 step0to1	2.59	0.34	1.07	(0.84, 1.16)	0.9	1.00	(0.38, 1.62)	0.1

Table D.2

Parameter Estimates and Fit: Multidimensional Model, Multidimensional Rubric

Parameter	Parameter estimate	Standard error	Unweighted Mean Square	Confidence Interval	T	Weighted Mean Square	Confidence Interval	T
Argument item1	-0.90	0.24	1.46	(0.73, 1.27)	3.0	1.60	(0.67, 1.33)	3.1
Argument item2	-0.73	0.16	1.20	(0.71, 1.29)	1.3	1.33	(0.68, 1.32)	1.9
Argument item3	-0.87	0.24	1.22	(0.75, 1.25)	1.7	1.21	(0.74, 1.26)	1.5
Argument item4	-1.21	0.23	1.50	(0.74, 1.26)	3.3	1.60	(0.71, 1.29)	3.5
Argument item5	-0.89	0.20	1.27	(0.76, 1.24)	2.1	1.32	(0.73, 1.27)	2.2
Argument item6	-0.88	0.19	1.22	(0.75, 1.25)	1.7	1.14	(0.75, 1.25)	1.1
Argument item7	-1.13	0.21	1.11	(0.75, 1.25)	0.8	1.30	(0.71, 1.29)	1.9
Argument item8	-1.84	0.25	1.39	(0.77, 1.23)	2.9	1.48	(0.73, 1.27)	3.1
Argument item9	-1.24	0.27	1.39	(0.74, 1.26)	2.6	1.46	(0.71, 1.29)	2.8
Argument item10	-1.31	0.21	1.38	(0.76, 1.24)	2.9	1.51	(0.72, 1.28)	3.1
Argument item11	-1.20	0.25	1.72	(0.74, 1.26)	4.5	1.29	(0.66, 1.34)	1.6
Argument item12	-0.90	0.26	1.03	(0.75, 1.25)	0.3	1.09	(0.72, 1.28)	0.7
Argument item13	-1.02	0.20	1.14	(0.75, 1.25)	1.1	1.28	(0.70, 1.30)	1.7
Argument item14	-0.96	0.17	1.40	(0.76, 1.24)	2.9	1.36	(0.75, 1.25)	2.5
Argument item15A	-1.21	0.21	1.03	(0.75, 1.25)	0.2	1.19	(0.70, 1.30)	1.2
Argument item15B	-0.89	0.28	1.03	(0.74, 1.26)	0.3	1.23	(0.70, 1.30)	1.4
Argument item16A	-0.87	0.15	1.12	(0.74, 1.26)	1.0	1.10	(0.73, 1.27)	0.7
Argument item16B	-0.31	0.13	1.40	(0.73, 1.27)	2.6	1.41	(0.73, 1.27)	2.7
Argument item17	-1.22	0.27	1.00	(0.76, 1.24)	0.1	1.32	(0.71, 1.29)	2.0
Argument item18	-1.88	0.28	1.49	(0.74, 1.26)	3.3	1.57	(0.70, 1.30)	3.2
Argument item19	-0.80	0.14	1.13	(0.84, 1.16)	1.6	1.32	(0.81, 1.19)	3.0
Argument item20	-1.09	0.16	1.00	(0.84, 1.16)	0.0	1.28	(0.79, 1.21)	2.4

Parameter	Parameter estimate	Standard error	Unweighted Mean Square	Confidence Interval	T	Weighted Mean Square	Confidence Interval	T
Argument rater1	-0.02	0.16	1.10	(0.85, 1.15)	1.3	1.11	(0.82, 1.18)	1.2
Argument rater2	-0.39	0.13	1.05	(0.85, 1.15)	0.7	1.08	(0.83, 1.17)	1.0
Argument rater3	-0.38	0.15	1.12	(0.85, 1.15)	1.5	1.19	(0.82, 1.18)	2.1
Argument rater4	0.22	0.09	1.13	(0.85, 1.15)	1.7	1.20	(0.84, 1.16)	2.3
Argument rater5	0.30	0.15	1.10	(0.85, 1.15)	1.3	1.01	(0.84, 1.16)	0.1
Argument rater6	0.20	0.12	0.93	(0.85, 1.15)	-0.8	0.94	(0.82, 1.18)	-0.6
Argument item1 x step0to1	0.19	0.57	1.32	(0.73, 1.27)	2.2	1.47	(0.78, 1.22)	3.7
Argument item1 x step1to2	-2.04	0.48	1.30	(0.73, 1.27)	2.1	1.43	(0.79, 1.21)	3.6
Argument item2 x step0to1	0.03	0.34	1.22	(0.71, 1.29)	1.4	1.29	(0.77, 1.23)	2.3
Argument item2 x step1to2	-1.48	0.32	1.39	(0.71, 1.29)	2.4	1.45	(0.79, 1.21)	3.7
Argument item3 x step0to1	-1.91	0.45	1.22	(0.75, 1.25)	1.7	1.25	(0.78, 1.22)	2.1
Argument item3 x step1to2	0.14	0.30	1.14	(0.75, 1.25)	1.1	1.17	(0.82, 1.18)	1.8
Argument item4 x step0to1	-0.72	0.46	1.54	(0.74, 1.26)	3.6	1.65	(0.80, 1.20)	5.2
Argument item4 x step1to2	-0.76	0.34	1.47	(0.74, 1.26)	3.2	1.57	(0.81, 1.19)	5.0
Argument item5 x step0to1	-0.40	0.39	0.95	(0.76, 1.24)	-0.4	1.00	(0.81, 1.19)	0.0
Argument item5 x step1to2	-0.39	0.33	0.79	(0.76, 1.24)	-1.8	0.94	(0.79, 1.21)	-0.5
Argument item6 x step0to1	-0.97	0.30	1.24	(0.75, 1.25)	1.8	1.29	(0.80, 1.20)	2.6
Argument item6 x step1to2	-0.32	0.25	1.16	(0.75, 1.25)	1.3	1.21	(0.82, 1.18)	2.2
Argument item7 x step0to1	-0.44	0.46	1.19	(0.75, 1.25)	1.4	1.28	(0.79, 1.21)	2.5
Argument item7 x step1to2	-1.40	0.38	1.16	(0.75, 1.25)	1.3	1.24	(0.81, 1.19)	2.3
Argument item8 x step0to1	-0.95	0.48	1.46	(0.77, 1.23)	3.4	1.54	(0.82, 1.18)	4.9
Argument item8 x step1to2	-0.67	0.34	1.39	(0.77, 1.23)	3.0	1.49	(0.83, 1.17)	4.9
Argument item9 x step0to1	-1.59	0.54	1.13	(0.74, 1.26)	1.0	1.16	(0.77, 1.23)	1.4
Argument item9 x step1to2	-0.82	0.36	1.08	(0.74, 1.26)	0.6	1.13	(0.80, 1.20)	1.2
Argument item10 x step0to1	-0.32	0.39	1.17	(0.76, 1.24)	1.4	1.22	(0.81, 1.19)	2.2

Parameter	Parameter estimate	Standard error	Unweighted Mean Square	Confidence Interval	T	Weighted Mean Square	Confidence Interval	T
Argument item10 x step1to2	-1.20	0.32	1.23	(0.76, 1.24)	1.8	1.31	(0.83, 1.17)	3.3
Argument item11 x step0to1	-0.85	0.50	1.03	(0.74, 1.26)	0.3	1.27	(0.77, 1.23)	2.2
Argument item11 x step1to2	-0.40	0.37	1.02	(0.74, 1.26)	0.2	1.35	(0.76, 1.24)	2.7
Argument item12 x step0to1	-0.67	0.65	0.70	(0.75, 1.25)	-2.7	0.94	(0.72, 1.28)	-0.4
Argument item12 x step1to2	-0.93	0.56	0.94	(0.75, 1.25)	-0.4	0.94	(0.82, 1.18)	-0.6
Argument item13 x step0to1	-0.89	0.44	0.95	(0.75, 1.25)	-0.3	1.20	(0.73, 1.27)	1.4
Argument item13 x step1to2	-1.43	0.37	1.12	(0.75, 1.25)	0.9	1.21	(0.80, 1.20)	1.9
Argument item14 x step0to1	0.07	0.36	1.14	(0.76, 1.24)	1.1	1.21	(0.81, 1.19)	2.2
Argument item14 x step1to2	-1.19	0.32	1.10	(0.76, 1.24)	0.8	1.23	(0.81, 1.19)	2.2
Argument item15A x step0to1	-1.02	0.41	1.17	(0.75, 1.25)	1.3	1.37	(0.76, 1.24)	2.8
Argument item15A x step1to2	-0.33	0.33	1.13	(0.75, 1.25)	1.1	1.20	(0.79, 1.21)	1.8
Argument item15B x step0to1	-2.44	0.52	0.98	(0.74, 1.26)	-0.1	1.26	(0.66, 1.34)	1.4
Argument item15B x step1to2	-0.48	0.32	1.04	(0.74, 1.26)	0.4	1.15	(0.77, 1.23)	1.3
Argument item16A x step0to1	-0.09	0.31	1.22	(0.74, 1.26)	1.6	1.29	(0.81, 1.19)	2.7
Argument item16A x step1to2	-0.37	0.29	1.14	(0.74, 1.26)	1.1	1.21	(0.79, 1.21)	1.9
Argument item16B x step0to1	-0.10	0.26	1.15	(0.73, 1.27)	1.1	1.15	(0.80, 1.20)	1.4
Argument item16B x step1to2	-0.68	0.26	1.01	(0.73, 1.27)	0.1	1.03	(0.78, 1.22)	0.3
Argument item17 x step0to1	-1.07	0.56	1.03	(0.76, 1.24)	0.3	1.26	(0.78, 1.22)	2.1
Argument item17 x step1to2	-0.96	0.39	1.04	(0.76, 1.24)	0.4	1.24	(0.79, 1.21)	2.2
Argument item18 x step0to1	-0.20	0.67	1.59	(0.74, 1.26)	3.9	1.70	(0.80, 1.20)	5.6
Argument item18 x step1to2	-1.72	0.53	1.51	(0.74, 1.26)	3.4	1.62	(0.80, 1.20)	5.2
Argument item19 x step0to1	-0.84	0.27	1.15	(0.84, 1.16)	1.7	1.28	(0.85, 1.15)	3.5
Argument item19 x step1to2	-0.45	0.24	1.14	(0.84, 1.16)	1.6	1.22	(0.89, 1.11)	3.6
Argument item20 x step0to1	-0.14	0.34	1.03	(0.84, 1.16)	0.4	1.11	(0.88, 1.12)	1.8
Argument item20 x step1to2	-1.39	0.29	0.99	(0.84, 1.16)	-0.2	1.05	(0.89, 1.11)	0.8

Parameter	Parameter estimate	Standard error	Unweighted Mean Square	Confidence Interval	T	Weighted Mean Square	Confidence Interval	T
Argument item1 x rater1	0.15	0.17	0.70	(0.85, 1.15)	-4.5	0.69	(0.81, 1.19)	-3.7
Argument item3 x rater2	0.00	0.19	0.72	(0.85, 1.15)	-4.2	0.66	(0.82, 1.18)	-4.1
Argument item7 x rater1	0.11	0.15	0.73	(0.85, 1.15)	-3.8	0.65	(0.79, 1.21)	-3.8
Argument item8 x rater4	0.04	0.17	0.66	(0.83, 1.17)	-4.3	0.64	(0.80, 1.20)	-3.9
Argument item9 x rater2	-0.05	0.22	0.61	(0.80, 1.20)	-4.5	0.60	(0.78, 1.22)	-4.2
Argument item10 x rater6	0.14	0.13	0.85	(0.80, 1.20)	-1.5	0.84	(0.78, 1.22)	-1.5
Argument item11 x rater2	0.14	0.19	0.73	(0.85, 1.15)	-3.8	0.73	(0.80, 1.20)	-2.9
Argument item12 x rater2	-0.36	0.17	0.80	(0.84, 1.16)	-2.7	0.73	(0.82, 1.18)	-3.2
Argument item13 x rater1	0.04	0.15	0.87	(0.85, 1.15)	-1.7	0.82	(0.82, 1.18)	-2.0
Argument item14 x rater4	-0.17	0.12	0.73	(0.80, 1.20)	-2.9	0.62	(0.77, 1.23)	-3.7
Argument item15A x rater3	0.17	0.16	0.73	(0.85, 1.15)	-3.8	0.71	(0.79, 1.21)	-3.0
Argument item15B x rater3	0.11	0.22	0.70	(0.85, 1.15)	-4.2	0.67	(0.80, 1.20)	-3.6
Argument item16A x rater4	0.32	0.11	0.37	(0.74, 1.26)	-6.4	0.31	(0.72, 1.28)	-6.8
Argument item17 x rater1	-0.29	0.20	0.69	(0.85, 1.15)	-4.5	0.67	(0.81, 1.19)	-3.9
Argument item18 x rater4	-0.16	0.19	0.84	(0.84, 1.16)	-2.1	0.73	(0.81, 1.19)	-3.1
Argument item19 x rater2	0.14	0.13	0.72	(0.84, 1.16)	-3.9	0.65	(0.80, 1.20)	-4.0
Argument item1 x step0to1 x rater1	0.15	0.57	0.67	(0.73, 1.27)	-2.8	0.65	(0.78, 1.22)	-3.5
Argument item1 x step1to2 x rater1	-0.25	0.48	0.70	(0.73, 1.27)	-2.4	0.70	(0.79, 1.21)	-3.1
Argument item2 x step0to1 x rater5	-0.09	0.34	0.68	(0.71, 1.29)	-2.4	0.68	(0.77, 1.23)	-3.1
Argument item2 x step1to2 x rater5	0.20	0.32	0.59	(0.71, 1.29)	-3.3	0.56	(0.79, 1.21)	-4.9
Argument item3 x step0to1 x rater2	0.00	0.45	0.79	(0.75, 1.25)	-1.7	0.69	(0.78, 1.22)	-3.0
Argument item3 x step1to2 x rater2	-0.08	0.29	0.83	(0.75, 1.25)	-1.4	0.80	(0.82, 1.18)	-2.3
Argument item4 x step0to1 x rater2	-0.64	0.46	0.58	(0.74, 1.26)	-3.7	0.42	(0.80, 1.20)	-7.2
Argument item4 x step1to2 x rater2	0.22	0.34	0.57	(0.74, 1.26)	-3.9	0.45	(0.81, 1.19)	-7.2
Argument item5 x step0to1 x rater1	-0.36	0.38	0.91	(0.76, 1.24)	-0.7	0.94	(0.81, 1.19)	-0.6

Parameter	Parameter estimate	Standard error	Unweighted Mean Square	Confidence Interval	T	Weighted Mean Square	Confidence Interval	T
Argument item5 x step1to2 x rater1	0.62	0.33	0.83	(0.76, 1.24)	-1.4	1.00	(0.79, 1.21)	0.0
Argument item6 x step0to1 x rater6	0.77	0.30	0.86	(0.75, 1.25)	-1.2	0.79	(0.80, 1.20)	-2.1
Argument item6 x step1to2 x rater6	-0.69	0.25	0.84	(0.75, 1.25)	-1.3	0.81	(0.82, 1.18)	-2.2
Argument item7 x step0to1 x rater1	-0.35	0.45	0.67	(0.75, 1.25)	-2.9	0.64	(0.79, 1.21)	-3.8
Argument item7 x step1to2 x rater1	0.39	0.38	0.72	(0.75, 1.25)	-2.3	0.69	(0.81, 1.19)	-3.7
Argument item8 x step0to1 x rater4	-0.22	0.48	0.71	(0.77, 1.23)	-2.7	0.55	(0.82, 1.18)	-5.7
Argument item8 x step1to2 x rater4	0.13	0.34	0.66	(0.77, 1.23)	-3.2	0.52	(0.83, 1.17)	-6.7
Argument item9 x step0to1 x rater2	0.62	0.54	0.78	(0.74, 1.26)	-1.8	0.77	(0.77, 1.23)	-2.2
Argument item9 x step1to2 x rater2	-0.21	0.36	0.82	(0.74, 1.26)	-1.4	0.81	(0.80, 1.20)	-2.0
Argument item10 x step0to1 x rater6	0.45	0.39	0.82	(0.76, 1.24)	-1.6	0.76	(0.81, 1.19)	-2.7
Argument item10 x step1to2 x rater6	-0.48	0.32	0.74	(0.76, 1.24)	-2.3	0.68	(0.83, 1.17)	-4.0
Argument item11 x step0to1 x rater2	-0.61	0.50	0.57	(0.74, 1.26)	-3.8	0.64	(0.77, 1.23)	-3.4
Argument item11 x step1to2 x rater2	-0.17	0.37	0.48	(0.74, 1.26)	-4.8	0.59	(0.76, 1.24)	-4.0
Argument item12 x step0to1 x rater2	1.43	0.65	0.72	(0.75, 1.25)	-2.4	0.98	(0.72, 1.28)	-0.1
Argument item12 x step1to2 x rater2	-0.56	0.56	1.01	(0.75, 1.25)	0.1	1.04	(0.82, 1.18)	0.4
Argument item13 x step0to1 x rater1	0.75	0.44	0.74	(0.75, 1.25)	-2.3	0.87	(0.73, 1.27)	-1.0
Argument item13 x step1to2 x rater1	-0.44	0.36	0.84	(0.75, 1.25)	-1.3	0.83	(0.80, 1.20)	-1.7
Argument item14 x step0to1 x rater4	-0.08	0.35	0.82	(0.76, 1.24)	-1.5	0.80	(0.81, 1.19)	-2.3
Argument item14 x step1to2 x rater4	-0.55	0.32	0.76	(0.76, 1.24)	-2.0	0.81	(0.81, 1.19)	-2.0
Argument item15A x step0to1 x rater3	0.42	0.41	0.63	(0.75, 1.25)	-3.3	0.65	(0.76, 1.24)	-3.3
Argument item15A x step1to2 x rater3	0.65	0.32	0.81	(0.75, 1.25)	-1.5	0.80	(0.79, 1.21)	-2.0
Argument item15B x step0to1 x rater3	0.40	0.51	0.51	(0.74, 1.26)	-4.5	0.73	(0.66, 1.34)	-1.7
Argument item15B x step1to2 x rater3	0.44	0.32	0.78	(0.74, 1.26)	-1.7	0.80	(0.77, 1.23)	-1.8
Argument item16A x step0to1 x rater4	0.82	0.31	0.67	(0.74, 1.26)	-2.8	0.60	(0.81, 1.19)	-4.7
Argument item16A x step1to2 x rater4	-1.17	0.29	0.80	(0.74, 1.26)	-1.6	0.78	(0.79, 1.21)	-2.2

Parameter	Parameter estimate	Standard error	Unweighted Mean Square	Confidence Interval	T	Weighted Mean Square	Confidence Interval	T
Argument item16B x step0to1 x rater4	0.45	0.26	0.89	(0.73, 1.27)	-0.8	0.84	(0.80, 1.20)	-1.6
Argument item16B x step1to2 x rater4	-0.70	0.26	1.00	(0.73, 1.27)	0.0	1.05	(0.78, 1.22)	0.4
Argument item17 x step0to1 x rater1	0.02	0.56	0.62	(0.76, 1.24)	-3.5	0.68	(0.78, 1.22)	-3.2
Argument item17 x step1to2 x rater1	0.66	0.39	0.64	(0.76, 1.24)	-3.3	0.70	(0.79, 1.21)	-3.2
Argument item18 x step0to1 x rater4	-0.02	0.67	0.56	(0.74, 1.26)	-4.0	0.40	(0.80, 1.20)	-7.6
Argument item18 x step1to2 x rater4	0.20	0.53	0.59	(0.74, 1.26)	-3.7	0.45	(0.80, 1.20)	-7.0
Argument item19 x step0to1 x rater2	0.22	0.27	0.85	(0.84, 1.16)	-1.9	0.76	(0.85, 1.15)	-3.5
Argument item19 x step1to2 x rater2	0.53	0.24	0.87	(0.84, 1.16)	-1.6	0.80	(0.89, 1.11)	-3.8
Argument item20 x step0to1 x rater3	0.73	0.33	0.87	(0.84, 1.16)	-1.7	0.85	(0.88, 1.12)	-2.6
Argument item20 x step1to2 x rater3	-0.41	0.29	0.92	(0.84, 1.16)	-1.1	0.92	(0.89, 1.11)	-1.4
SPQ item1	-1.17	0.23	1.34	(0.76, 1.24)	2.5	1.08	(0.80, 1.20)	0.8
SPQ item2	-0.44	0.24	1.06	(0.72, 1.28)	0.5	1.00	(0.82, 1.18)	0.0
SPQ item3	-0.56	0.15	1.20	(0.76, 1.24)	1.6	1.09	(0.79, 1.21)	0.8
SPQ item4	-0.27	0.15	1.34	(0.74, 1.26)	2.3	1.14	(0.78, 1.22)	1.3
SPQ item5	-0.79	0.21	1.13	(0.76, 1.24)	1.1	1.10	(0.83, 1.17)	1.1
SPQ item6	-1.49	0.30	1.46	(0.70, 1.30)	2.7	1.08	(0.69, 1.31)	0.5
SPQ item7	-1.54	0.24	1.00	(0.76, 1.24)	0.0	1.02	(0.74, 1.26)	0.2
SPQ item8	-1.59	0.23	0.77	(0.77, 1.23)	-2.1	0.91	(0.77, 1.23)	-0.8
SPQ item9	-2.40	0.31	1.01	(0.75, 1.25)	0.2	1.06	(0.59, 1.41)	0.4
SPQ item10	-1.91	0.26	1.15	(0.77, 1.23)	1.2	1.00	(0.70, 1.30)	0.0
SPQ item11	-1.90	0.27	0.87	(0.75, 1.25)	-1.0	0.97	(0.69, 1.31)	-0.1
SPQ item12	-0.42	0.25	0.88	(0.71, 1.29)	-0.8	0.94	(0.83, 1.17)	-0.7
SPQ item13	0.73	0.19	0.93	(0.76, 1.24)	-0.6	0.93	(0.78, 1.22)	-0.6
SPQ item14	2.78	0.33	0.65	(0.76, 1.25)	-3.2	0.92	(0.56, 1.44)	-0.3
SPQ item15A	0.55	0.18	1.05	(0.68, 1.32)	0.3	1.11	(0.72, 1.28)	0.7

Parameter	Parameter estimate	Standard error	Unweighted Mean Square	Confidence Interval	T	Weighted Mean Square	Confidence Interval	T
SPQ item15B	1.66	0.23	0.73	(0.70, 1.30)	-1.9	0.87	(0.58, 1.42)	-0.6
SPQ item16A	-0.19	0.18	0.80	(0.71, 1.29)	-1.4	0.86	(0.75, 1.25)	-1.1
SPQ item16B	0.43	0.17	0.72	(0.70, 1.30)	-2.0	0.84	(0.74, 1.26)	-1.3
SPQ item17	-0.19	0.21	0.98	(0.76, 1.24)	-0.1	0.96	(0.86, 1.14)	-0.6
SPQ item18	-0.41	0.22	1.05	(0.74, 1.26)	0.4	1.04	(0.85, 1.15)	0.5
SPQ item19	-1.14	0.12	0.88	(0.84, 1.16)	-1.5	0.91	(0.85, 1.15)	-1.2
SPQ item20	-1.02	0.14	0.87	(0.84, 1.16)	-1.6	0.93	(0.88, 1.12)	-1.3
SPQ item3 x step 0to1	0.26	0.23	0.83	(0.76, 1.24)	-1.4	0.94	(0.76, 1.24)	-0.4
SPQ item4 x step0to1	0.47	0.25	0.98	(0.74, 1.26)	-0.1	0.99	(0.70, 1.30)	0.0
SPQ item13 x step0to1	-0.89	0.30	0.91	(0.76, 1.24)	-0.7	0.94	(0.78, 1.22)	-0.6
SPQ item13 x step1to2	-1.79	0.29	0.94	(0.76, 1.24)	-0.4	0.98	(0.85, 1.15)	-0.2
SPQ item15A x step0to1	0.62	0.32	1.18	(0.68, 1.32)	1.1	1.06	(0.59, 1.41)	0.3
SPQ item15B x step0to1	0.95	0.42	0.81	(0.70, 1.30)	-1.3	0.96	(0.38, 1.62)	0.0
SPQ item16A x step0to1	-0.18	0.24	0.96	(0.71, 1.29)	-0.2	0.99	(0.81, 1.19)	-0.1
SPQ item16B x step0to1	0.77	0.32	0.92	(0.70, 1.30)	-0.5	0.98	(0.57, 1.43)	0.0
SPQ item19 x step0to1	-0.76	0.15	0.88	(0.84, 1.16)	-1.5	0.91	(0.93, 1.07)	-2.4
Matter item1	-0.43	0.13	1.11	(0.74, 1.26)	0.9	1.04	(0.80, 1.20)	0.4
Matter item2	-0.99	0.20	0.92	(0.69, 1.31)	-0.5	1.02	(0.56, 1.44)	0.1
Matter item3	0.34	0.13	0.91	(0.74, 1.26)	-0.6	0.97	(0.80, 1.20)	-0.3
Matter item4	0.32	0.13	1.08	(0.73, 1.27)	0.6	1.09	(0.81, 1.19)	0.9
Matter item5	-0.22	0.12	0.84	(0.76, 1.24)	-1.4	0.94	(0.82, 1.18)	-0.7
Matter item6	-0.22	0.12	0.87	(0.75, 1.25)	-1.0	0.94	(0.83, 1.17)	-0.7
Matter item7	0.58	0.12	1.48	(0.75, 1.25)	3.4	1.33	(0.80, 1.20)	3.0
Matter item8	0.21	0.11	1.12	(0.77, 1.23)	1.0	1.10	(0.82, 1.18)	1.1
Matter item9	-0.21	0.12	1.21	(0.75, 1.25)	1.6	1.16	(0.82, 1.18)	1.7

Parameter	Parameter estimate	Standard error	Unweighted Mean Square	Confidence Interval	T	Weighted Mean Square	Confidence Interval	T
Matter item10	0.00	0.11	1.26	(0.76, 1.24)	2.0	1.20	(0.84, 1.16)	2.3
Matter item11	0.35	0.12	1.15	(0.75, 1.25)	1.1	1.15	(0.79, 1.21)	1.4
Matter item12	0.02	0.12	1.00	(0.74, 1.26)	0.0	1.04	(0.83, 1.17)	0.5
Matter item13	1.64	0.25	0.86	(0.76, 1.24)	-1.2	0.91	(0.75, 1.25)	-0.7
Matter item14	0.75	0.16	0.99	(0.76, 1.24)	0.0	0.97	(0.80, 1.20)	-0.2
Matter item15	-0.02	0.12	1.03	(0.74, 1.26)	0.3	1.02	(0.82, 1.18)	0.2
Matter item16	-0.97	0.19	0.67	(0.69, 1.31)	-2.4	0.97	(0.58, 1.42)	-0.1
Matter item17	0.21	0.26	1.07	(0.76, 1.24)	0.6	0.98	(0.83, 1.17)	-0.2
Matter item18	-0.84	0.14	0.70	(0.74, 1.26)	-2.5	0.89	(0.74, 1.26)	-0.8
Matter item19	-0.20	0.08	0.90	(0.85, 1.15)	-1.3	0.94	(0.89, 1.11)	-1.1
Matter item20	-0.45	0.08	1.06	(0.84, 1.16)	0.8	1.02	(0.88, 1.12)	0.3
Matter item1 step0to1	1.64	0.37	1.12	(0.74, 1.26)	0.9	1.02	(0.39, 1.61)	0.2
Matter item2 step0to1	2.36	0.73	1.08	(0.69, 1.31)	0.5	0.99	(0.00, 2.31)	0.2
Matter item3 step0to1	0.77	0.26	0.98	(0.74, 1.26)	-0.1	1.00	(0.66, 1.34)	0.1
Matter item4 step0to1	1.19	0.31	1.05	(0.73, 1.27)	0.4	1.00	(0.54, 1.46)	0.1
Matter item5 step0to1	1.02	0.26	0.92	(0.76, 1.24)	-0.7	0.98	(0.62, 1.38)	0.0
Matter item6 step0to1	1.26	0.29	0.91	(0.75, 1.25)	-0.7	0.99	(0.55, 1.45)	0.0
Matter item7 step0to1	2.49	0.51	1.02	(0.75, 1.25)	0.2	1.01	(0.07, 1.93)	0.2
Matter item8 step0to1	2.39	0.46	1.42	(0.77, 1.23)	3.2	1.02	(0.17, 1.83)	0.2
Matter item9 step0to1	1.22	0.29	0.99	(0.75, 1.25)	0.0	1.00	(0.55, 1.45)	0.1
Matter item10 step0to1	4.03	1.00	0.87	(0.76, 1.24)	-1.1	1.00	(0.00, 2.93)	0.3
Matter item11 step0to1	2.72	0.59	0.86	(0.75, 1.25)	-1.1	1.00	(0.00, 2.09)	0.2
Matter item 12 step0to1	3.88	1.01	1.06	(0.74, 1.26)	0.5	1.00	(0.00, 2.93)	0.3
Matter item13 step0to1	-0.96	0.28	0.95	(0.76, 1.24)	-0.4	0.99	(0.85, 1.15)	-0.1
Matter item14 step0to1	-0.91	0.19	0.95	(0.76, 1.24)	-0.4	0.96	(0.94, 1.06)	-1.3

Parameter	Parameter estimate	Standard error	Unweighted Mean Square	Confidence Interval	T	Weighted Mean Square	Confidence Interval	T
Matter item15 step0to1	2.22	0.46	1.25	(0.74, 1.26)	1.8	1.01	(0.19, 1.81)	0.2
Matter item16 step0to1	1.68	0.53	0.72	(0.69, 1.31)	-1.9	0.98	(0.10, 1.90)	0.1
Matter item17 step0to1	-1.03	0.44	1.06	(0.76, 1.24)	0.5	0.99	(0.86, 1.14)	-0.1
Matter item18 step0to1	2.55	0.59	0.88	(0.74, 1.26)	-0.9	0.99	(0.00, 2.08)	0.2
Matter item19 step0to1	1.25	0.18	1.01	(0.85, 1.15)	0.1	1.00	(0.72, 1.28)	0.0
Matter item20 step0to1	2.58	0.34	1.10	(0.84, 1.16)	1.3	1.00	(0.38, 1.62)	0.1

Table D.3

Parameter Estimates and Fit: Unidimensional Model, Holistic Rubric

Parameter	Parameter estimate	Standard error	Unweighted Mean Square	Confidence Interval	T	Weighted Mean Square	Confidence Interval	T
item1	-0.65	0.15	1.49	(0.76, 1.24)	3.5	1.49	(0.77, 1.23)	3.7
item2	-0.25	0.11	1.59	(0.74, 1.26)	3.8	1.62	(0.73, 1.27)	3.9
item3	-0.44	0.14	1.33	(0.76, 1.24)	2.4	1.45	(0.77, 1.23)	3.4
item4	-0.14	0.12	1.43	(0.74, 1.26)	2.9	1.49	(0.75, 1.25)	3.3
item5	-0.12	0.11	1.23	(0.76, 1.24)	1.8	1.38	(0.76, 1.24)	2.8
item6	-0.37	0.12	1.26	(0.76, 1.24)	2.0	1.29	(0.76, 1.24)	2.2
item7	-0.72	0.11	1.71	(0.77, 1.23)	5.0	1.59	(0.74, 1.26)	3.8
item8	-0.34	0.10	1.12	(0.77, 1.23)	1.0	1.21	(0.76, 1.24)	1.6
item9	-0.94	0.17	1.51	(0.75, 1.25)	3.5	1.65	(0.76, 1.24)	4.5
item10	-0.96	0.20	1.54	(0.77, 1.23)	3.9	1.69	(0.77, 1.23)	5.0
item11	-0.30	0.12	1.28	(0.75, 1.25)	2.1	1.32	(0.77, 1.23)	2.4
item12	-0.07	0.11	1.41	(0.76, 1.24)	3.0	1.55	(0.76, 1.24)	3.9
item13	1.42	0.17	1.35	(0.76, 1.24)	2.6	1.57	(0.56, 1.44)	2.2
item14	1.04	0.13	1.28	(0.77, 1.23)	2.2	1.56	(0.65, 1.35)	2.7
item15	0.57	0.15	0.99	(0.75, 1.25)	0.0	1.08	(0.72, 1.28)	0.6
item16	-0.02	0.12	1.32	(0.73, 1.27)	2.2	1.26	(0.74, 1.26)	1.9
item17	-0.76	0.13	1.12	(0.76, 1.24)	1.0	1.21	(0.75, 1.25)	1.6
item18	-0.97	0.14	1.05	(0.74, 1.26)	0.4	1.14	(0.72, 1.28)	1.0
item19	-0.34	0.09	1.20	(0.85, 1.15)	2.5	1.29	(0.85, 1.15)	3.6
item20	-0.90	0.16	1.34	(0.84, 1.16)	3.9	1.40	(0.83, 1.17)	4.2
rater2	-0.04	0.09	0.89	(0.85, 1.15)	-1.5	0.92	(0.85, 1.15)	-1.1
rater4	0.06	0.07	0.81	(0.86, 1.14)	-2.7	0.82	(0.85, 1.15)	-2.5

Parameter	Parameter estimate	Standard error	Unweighted Mean Square	Confidence Interval	T	Weighted Mean Square	Confidence Interval	T
rater5	0.34	0.05	1.05	(0.86, 1.14)	0.7	1.10	(0.85, 1.15)	1.2
rater6	-0.46	0.08	0.78	(0.86, 1.14)	-3.2	0.79	(0.85, 1.15)	-2.9
item1 x step0to1	0.10	0.34	1.11	(0.76, 1.24)	0.9	1.13	(0.79, 1.21)	1.2
item1 x step1to2	-0.24	0.30	1.27	(0.76, 1.24)	2.1	1.36	(0.75, 1.25)	2.6
item2 x step0to1	0.80	0.29	1.42	(0.74, 1.26)	2.9	1.49	(0.79, 1.21)	4.0
item2 x step1to2	-0.81	0.31	1.58	(0.74, 1.26)	3.8	1.58	(0.75, 1.25)	3.9
item3 x step0to1	-1.03	0.24	1.15	(0.76, 1.24)	1.2	1.22	(0.82, 1.18)	2.2
item3 x step1to2	0.34	0.22	1.17	(0.76, 1.24)	1.3	1.19	(0.78, 1.22)	1.6
item4 x step0to1	-0.75	0.21	1.13	(0.74, 1.26)	1.0	1.16	(0.81, 1.19)	1.5
item4 x step1to2	-0.09	0.23	0.86	(0.74, 1.26)	-1.1	0.78	(0.80, 1.20)	-2.3
item5 x step0to1	-0.78	0.20	1.19	(0.76, 1.24)	1.6	1.22	(0.82, 1.18)	2.3
item5 x step1to2	0.22	0.21	1.22	(0.76, 1.24)	1.8	1.27	(0.79, 1.21)	2.3
item6 x step0to1	-0.69	0.23	1.11	(0.76, 1.24)	0.9	1.09	(0.81, 1.19)	1.0
item6 x step1to2	0.20	0.23	1.08	(0.76, 1.24)	0.6	1.07	(0.77, 1.23)	0.6
item7 x step0to1	1.35	0.53	1.22	(0.77, 1.23)	1.8	1.16	(0.80, 1.20)	1.6
item7 x step1to2	-1.21	0.54	1.10	(0.77, 1.23)	0.8	1.17	(0.76, 1.24)	1.3
item8 x step0to1	1.59	0.52	1.12	(0.77, 1.23)	1.0	1.15	(0.83, 1.17)	1.6
item8 x step1to2	-1.87	0.54	1.08	(0.77, 1.23)	0.7	1.15	(0.82, 1.18)	1.7
item9 x step0to1	-1.81	0.33	1.14	(0.75, 1.25)	1.1	1.22	(0.79, 1.21)	1.9
item9 x step1to2	0.95	0.26	1.08	(0.75, 1.25)	0.7	1.13	(0.79, 1.21)	1.2
item10 x step0to1	-2.16	0.40	1.26	(0.77, 1.23)	2.0	1.29	(0.78, 1.22)	2.4
item10 x step1to2	1.02	0.27	0.82	(0.77, 1.23)	-1.6	0.78	(0.80, 1.20)	-2.3
item11 x step0to1	-0.06	0.25	1.14	(0.75, 1.25)	1.1	1.23	(0.80, 1.20)	2.2
item11 x step1to2	-0.67	0.26	1.22	(0.75, 1.25)	1.7	1.38	(0.78, 1.22)	3.1
item12 x step0to1	2.21	0.57	1.28	(0.76, 1.24)	2.1	1.41	(0.80, 1.20)	3.6

Parameter	Parameter estimate	Standard error	Unweighted Mean Square	Confidence Interval	T	Weighted Mean Square	Confidence Interval	T
item12 x step1to2	-2.74	0.58	1.24	(0.76, 1.24)	1.8	1.37	(0.79, 1.21)	3.1
item13 x step0to1	0.80	0.34	1.10	(0.76, 1.24)	0.8	1.26	(0.69, 1.31)	1.6
item13 x step1to2	0.19	0.57	0.55	(0.76, 1.24)	-4.3	0.95	(0.18, 1.82)	0.0
item14 x step0to1	0.92	0.27	1.22	(0.77, 1.23)	1.8	1.27	(0.72, 1.28)	1.8
item14 x step1to2	0.87	0.61	1.16	(0.77, 1.23)	1.3	1.01	(0.25, 1.75)	0.1
item15 x step0to1	-1.73	0.22	0.93	(0.75, 1.25)	-0.5	1.08	(0.74, 1.26)	0.6
item15 x step1to2	0.69	0.26	1.07	(0.75, 1.25)	0.6	1.18	(0.78, 1.22)	1.5
item16 x step0to1	-0.88	0.21	1.24	(0.73, 1.27)	1.7	1.27	(0.79, 1.21)	2.3
item16 x step1to2	0.16	0.22	1.21	(0.73, 1.27)	1.5	1.25	(0.79, 1.21)	2.2
item17 x step0to1	-0.75	0.28	1.07	(0.76, 1.24)	0.6	1.15	(0.80, 1.20)	1.4
item17 x step1to2	-0.33	0.26	1.02	(0.76, 1.24)	0.2	1.07	(0.83, 1.17)	0.8
item18 x step0to1	-0.68	0.30	1.15	(0.74, 1.26)	1.1	1.21	(0.79, 1.21)	1.8
item18 x step1to2	-0.49	0.27	1.08	(0.74, 1.26)	0.6	1.10	(0.81, 1.19)	1.0
item19 x step0to1	-0.48	0.17	1.12	(0.85, 1.15)	1.5	1.15	(0.90, 1.10)	2.9
item19 x step1to2	0.07	0.19	1.03	(0.85, 1.15)	0.5	1.10	(0.86, 1.14)	1.4
item20 x step0to1	0.62	0.27	1.12	(0.84, 1.16)	1.5	1.21	(0.88, 1.12)	3.3
item20 x step1to2	-0.55	0.28	1.15	(0.84, 1.16)	1.9	1.25	(0.86, 1.14)	3.4
item2 x rater5	-0.21	0.08	0.71	(0.83, 1.17)	-3.5	0.46	(0.82, 1.18)	-7.5
item3 x rater2	0.28	0.09	0.69	(0.80, 1.20)	-3.4	0.58	(0.78, 1.22)	-4.4
item 4 x rater2	-0.25	0.08	0.64	(0.83, 1.17)	-4.7	0.58	(0.80, 1.20)	-4.9
item5xrater6	0.00	0.09	0.66	(0.83, 1.17)	-4.4	0.54	(0.83, 1.17)	-6.4
item6xrater6	-0.05	0.10	0.69	(0.80, 1.20)	-3.3	0.59	(0.79, 1.21)	-4.4
item7 x rater4	-0.14	0.08	0.88	(0.85, 1.15)	-1.7	0.62	(0.82, 1.18)	-4.8
item8 x rater4	-0.23	0.08	0.75	(0.85, 1.15)	-3.5	0.61	(0.82, 1.18)	-5.0
item10 x rater6	-0.11	0.17	0.70	(0.83, 1.17)	-3.8	0.56	(0.83, 1.17)	-6.2

Parameter	Parameter estimate	Standard error	Unweighted Mean Square	Confidence Interval	T	Weighted Mean Square	Confidence Interval	T
item12 x rater2	-0.03	0.08	0.76	(0.83, 1.17)	-2.9	0.67	(0.83, 1.17)	-4.3
item13 x rater4	0.13	0.13	0.73	(0.85, 1.15)	-3.9	0.60	(0.83, 1.17)	-5.5
item14 x rater4	0.17	0.09	0.78	(0.85, 1.15)	-3.2	0.58	(0.83, 1.17)	-5.8
item15 x rater4	0.22	0.11	0.85	(0.85, 1.15)	-2.0	0.78	(0.83, 1.17)	-2.8
item16 x rater4	-0.20	0.11	1.07	(0.85, 1.15)	0.9	1.16	(0.84, 1.16)	1.9
item17 x rater4	-0.09	0.10	0.76	(0.85, 1.15)	-3.4	0.68	(0.84, 1.16)	-4.4
item18 x rater4	0.12	0.11	0.92	(0.85, 1.15)	-1.1	0.68	(0.83, 1.17)	-4.1
item19 x rater5	-0.05	0.07	0.80	(0.85, 1.15)	-2.8	0.68	(0.84, 1.16)	-4.2
item1 x step0to1 x rater5	1.18	0.33	0.84	(0.76, 1.24)	-1.3	0.75	(0.79, 1.21)	-2.6
item1 x step1to2 x rater5	-0.82	0.30	0.70	(0.76, 1.24)	-2.8	0.62	(0.75, 1.25)	-3.5
item2 x step0to1 x rater5	0.12	0.29	0.61	(0.74, 1.26)	-3.4	0.51	(0.79, 1.21)	-5.5
item2 x step1to2 x rater5	-0.08	0.31	0.65	(0.74, 1.26)	-3.0	0.43	(0.75, 1.25)	-5.8
item3 x step0to1 x rater2	0.40	0.24	0.71	(0.76, 1.24)	-2.6	0.65	(0.82, 1.18)	-4.3
item3 x step1to2 x rater2	-0.22	0.22	0.95	(0.76, 1.24)	-0.4	0.85	(0.78, 1.22)	-1.4
item4 x step0to1 x rater2	0.53	0.21	0.86	(0.74, 1.26)	-1.0	0.83	(0.81, 1.19)	-1.8
item4 x step1to2 x rater2	-0.67	0.23	1.22	(0.74, 1.26)	1.6	1.24	(0.80, 1.20)	2.2
item5 x step0to1 x rater6	-0.44	0.20	0.82	(0.76, 1.24)	-1.5	0.77	(0.82, 1.18)	-2.7
item5 x step1to2 x rater6	0.26	0.21	0.80	(0.76, 1.24)	-1.8	0.75	(0.79, 1.21)	-2.5
item6 x step0to1 x rater6	-0.63	0.23	0.92	(0.76, 1.24)	-0.6	0.85	(0.81, 1.19)	-1.6
item6 x step1to2 x rater6	0.74	0.23	0.94	(0.76, 1.24)	-0.4	0.89	(0.77, 1.23)	-0.9
item7 x step0to1 x rater4	-1.46	0.53	0.95	(0.77, 1.23)	-0.4	0.89	(0.80, 1.20)	-1.1
item7 x step1to2 x rater4	1.18	0.54	0.77	(0.77, 1.23)	-2.1	0.80	(0.76, 1.24)	-1.7
item8 x step0to1 x rater4	-1.51	0.52	0.80	(0.77, 1.23)	-1.8	0.74	(0.83, 1.17)	-3.2
item8 x step1to2 x rater4	0.92	0.54	0.79	(0.77, 1.23)	-1.9	0.76	(0.82, 1.18)	-2.8
item9 x step0to1 x rater6	0.26	0.33	0.82	(0.75, 1.25)	-1.5	0.77	(0.79, 1.21)	-2.3

Parameter	Parameter estimate	Standard error	Unweighted Mean Square	Confidence Interval	T	Weighted Mean Square	Confidence Interval	T
item9 x step1to2 x rater6	0.12	0.26	0.86	(0.75, 1.25)	-1.1	0.85	(0.79, 1.21)	-1.4
item10 x step0to1 x rater6	0.15	0.39	0.86	(0.77, 1.23)	-1.2	0.77	(0.78, 1.22)	-2.2
item10 x step1to2 x rater6	0.15	0.27	1.14	(0.77, 1.23)	1.1	1.19	(0.80, 1.20)	1.8
item11 x step0to1 x rater2	-1.14	0.25	0.73	(0.75, 1.25)	-2.3	0.67	(0.80, 1.20)	-3.6
item11 x step1to2 x rater2	0.53	0.26	0.70	(0.75, 1.25)	-2.6	0.58	(0.78, 1.22)	-4.4
item12 x step0to1 x rater2	0.09	0.57	0.70	(0.76, 1.24)	-2.6	0.62	(0.80, 1.20)	-4.3
item12 x step1to2 x rater2	-0.43	0.58	0.71	(0.76, 1.24)	-2.5	0.63	(0.79, 1.21)	-3.9
item13 x step0to1 x rater4	-0.73	0.34	0.61	(0.76, 1.24)	-3.7	0.70	(0.69, 1.31)	-2.1
item13 x step1to2 x rater4	1.15	0.57	0.60	(0.76, 1.24)	-3.8	1.12	(0.18, 1.82)	0.4
item14 x step0to1 x rater4	-0.58	0.27	0.75	(0.77, 1.23)	-2.3	0.71	(0.72, 1.28)	-2.2
item14 x step1to2 x rater4	-0.10	0.61	1.17	(0.77, 1.23)	1.4	1.04	(0.25, 1.75)	0.2
item15 x step0to1 x rater4	1.14	0.21	0.78	(0.75, 1.25)	-1.9	0.82	(0.74, 1.26)	-1.4
item15 x step1to2 x rater4	-0.05	0.26	0.78	(0.75, 1.25)	-1.9	0.74	(0.78, 1.22)	-2.5
item16 x step0to1 x rater4	0.11	0.21	0.70	(0.73, 1.27)	-2.4	0.63	(0.79, 1.21)	-4.0
item16 x step1to2 x rater4	0.49	0.22	0.74	(0.73, 1.27)	-2.0	0.73	(0.79, 1.21)	-2.7
item17 x step0to1 x rater4	0.88	0.28	0.77	(0.76, 1.24)	-2.0	0.75	(0.80, 1.20)	-2.6
item17 x step1to2 x rater4	-0.33	0.26	0.87	(0.76, 1.24)	-1.1	0.86	(0.83, 1.17)	-1.6
item18 x step0to1 x rater4	0.89	0.30	0.80	(0.74, 1.26)	-1.6	0.77	(0.79, 1.21)	-2.3
item18 x step1to2 x rater4	-0.34	0.27	0.87	(0.74, 1.26)	-1.0	0.86	(0.81, 1.19)	-1.5
item19 x step0to1 x rater5	0.18	0.16	0.99	(0.85, 1.15)	-0.1	0.95	(0.90, 1.10)	-1.0
item19 x step1to2 x rater5	-0.16	0.19	0.88	(0.85, 1.15)	-1.6	0.90	(0.86, 1.14)	-1.4
item20 x step0to1 x rater4	0.72	0.27	0.89	(0.84, 1.16)	-1.4	0.80	(0.88, 1.12)	-3.4
item20 x step1to2 x rater4	-0.85	0.28	0.88	(0.84, 1.16)	-1.5	0.76	(0.86, 1.14)	-3.7

Appendix E

Results of DIF Analysis

Table E.1

Parameter Estimates by Grade Level – Structure and Properties of Matter

Item	Grade	Estimate	Standard error
1 1A - Ana's ..	--	-0.348	0.118
2 1B - Ana's ..	--	-1.079	0.189
3 2A - Beth's..	--	0.454	0.122
4 2B - Beth's..	--	0.261	0.117
5 3A - Romita..	--	-0.151	0.110
6 3B - Romita..	--	-0.191	0.110
7 4A - Kevin'..	--	0.508	0.110
8 4B - Kevin'..	--	0.275	0.102
9 5A - Brass ..	--	-0.159	0.111
10 5B - Brass ..	--	0.003	0.101
11 6A - Carol'..	--	0.349	0.112
12 6B - Carol'..	--	0.109	0.111
13 7A - Box of..	--	1.640	0.170
14 7B - Box of..	--	0.832	0.145
15 8A - Grain ..	--	-0.011	0.109
16 8B - Grain ..	--	-0.867	0.179
17 9A - Nate's..	--	-0.220	0.114
18 9B - Nate's..	--	-0.774	0.127
19 10 - Drops ..	--	-0.125	0.069
20 11 - Ken's ..	--	-0.379	0.070
--	4	0.116[I]	0.024
--	5	-0.116*[I]	0.024
1 1A - Ana's 1	4	0.054	0.118
2 1B - Ana's 1	4	-0.136	0.189
3 2A - Beth's1	4	0.134	0.122
4 2B - Beth's1	4	-0.078	0.117
5 3A - Romita1	4	0.057	0.110
6 3B - Romita1	4	-0.103	0.110
7 4A - Kevin'1	4	-0.063	0.110
8 4B - Kevin'1	4	0.085	0.102
9 5A - Brass 1	4	0.038	0.111
10 5B - Brass 1	4	-0.263[I]	0.101

Item	Grade	Estimate	Standard error
11 6A - Carol'1	4	-0.163	0.112
12 6B - Carol'1	4	0.166	0.111
13 7A - Box of1	4	-0.213	0.170
14 7B - Box of1	4	0.112	0.145
15 8A - Grain 1	4	-0.162	0.109
16 8B - Grain 1	4	0.100	0.179
17 9A - Nate's1	4	0.246[1]	0.114
18 9B - Nate's1	4	-0.092	0.127
19 10 - Drops 1	4	0.065	0.069
20 11 - Ken's 1	4	0.019	0.070
1 1A - Ana's 2	5	-0.054*	0.118
2 1B - Ana's 2	5	0.136*	0.189
3 2A - Beth's2	5	-0.134*	0.122
4 2B - Beth's2	5	0.078*	0.117
5 3A - Romita2	5	-0.057*	0.110
6 3B - Romita2	5	0.103*	0.110
7 4A - Kevin'2	5	0.063*	0.110
8 4B - Kevin'2	5	-0.085*	0.102
9 5A - Brass 2	5	-0.038*	0.111
10 5B - Brass 2	5	0.263*[1]	0.101
11 6A - Carol'2	5	0.163*	0.112
12 6B - Carol'2	5	-0.166*	0.111
13 7A - Box of2	5	0.213*	0.170
14 7B - Box of2	5	-0.112*	0.145
15 8A - Grain 2	5	0.162*	0.109
16 8B - Grain 2	5	-0.100*	0.179
17 9A - Nate's2	5	-0.246*[1]	0.114
18 9B - Nate's2	5	0.092*	0.127
19 10 - Drops 2	5	-0.065*	0.069
20 11 - Ken's 2	5	-0.019*	0.070

* Indicates that parameter is constrained

[1] Indicates that difference between Grade 4 and Grade 5 parameters is statistically significant.

Table E.2

Parameter Estimates by Grade Level – Scale, Proportion, and Quantity

Item		Grade	Estimate	Standard error
1	1A - Ana's ..	--	-1.085	0.216
2	1B - Ana's ..	--	-0.475	0.233
3	2A - Beth's..	--	-0.559	0.130
4	2B - Beth's..	--	-0.303	0.135
5	3A - Romita..	--	-0.789	0.204
6	3B - Romita..	--	-1.458	0.293
7	4A - Kevin'..	--	-1.538	0.236
8	4B - Kevin'..	--	-1.484	0.220
9	5A - Brass ..	--	-2.290	0.301
10	5B - Brass ..	--	-1.951	0.249
11	6A - Carol'..	--	-1.803	0.258
12	6B - Carol'..	--	-0.351	0.241
13	7A - Box of..	--	0.839	0.121
14	7B - Box of..	--	3.294	0.351
15	8A.1 - Grai..	--	0.585	0.163
16	8A.2 - Grai..	--	1.769	0.206
17	8B.1 - Grai..	--	-0.179	0.161
18	8B.2 - Grai..	--	0.280	0.153
19	9A - Nate's..	--	-0.184	0.199
20	9B - Nate's..	--	-0.389	0.214
21	10 - Drops ..	--	-1.206	0.105
22	11 - Ken's ..	--	-1.053	0.139
--		4	0.116[1]	0.038
--		5	-0.116*[1]	0.038
1	1A - Ana's 1	4	0.257	0.216
2	1B - Ana's 1	4	-0.095	0.233
3	2A - Beth's1	4	-0.287[1]	0.130
4	2B - Beth's1	4	-0.059	0.135
5	3A - Romita1	4	-0.233	0.204
6	3B - Romita1	4	0.262	0.293
7	4A - Kevin'1	4	0.165	0.236
8	4B - Kevin'1	4	0.080	0.220
9	5A - Brass 1	4	0.389	0.301
10	5B - Brass 1	4	-0.236	0.249
11	6A - Carol'1	4	0.298	0.258
12	6B - Carol'1	4	-0.105	0.241
13	7A - Box of1	4	-0.234	0.121

Item	Grade	Estimate	Standard error
14 7B - Box of1	4	0.915[1]	0.351
15 8A.1 - Grai1	4	-0.078	0.163
16 8A.2 - Grai1	4	0.291	0.206
17 8B.1 - Grai1	4	0.148	0.161
18 8B.2 - Grai1	4	-0.078	0.153
19 9A - Nate's1	4	-0.148	0.199
20 9B - Nate's1	4	-0.058	0.214
21 10 - Drops 1	4	0.203	0.105
22 11 - Ken's 1	4	0.040	0.139
1 1A - Ana's 2	5	-0.257*	0.216
2 1B - Ana's 2	5	0.095*	0.233
3 2A - Beth's2	5	0.287*[1]	0.130
4 2B - Beth's2	5	0.059*	0.135
5 3A - Romita2	5	0.233*	0.204
6 3B - Romita2	5	-0.262*	0.293
7 4A - Kevin'2	5	-0.165*	0.236
8 4B - Kevin'2	5	-0.080*	0.220
9 5A - Brass 2	5	-0.389*	0.301
10 5B - Brass 2	5	0.236*	0.249
11 6A - Carol'2	5	-0.298*	0.258
12 6B - Carol'2	5	0.105*	0.241
13 7A - Box of2	5	0.234*	0.121
14 7B - Box of2	5	-0.915*[1]	0.351
15 8A.1 - Grai2	5	0.078*	0.163
16 8A.2 - Grai2	5	-0.291*	0.206
17 8B.1 - Grai2	5	-0.148*	0.161
18 8B.2 - Grai2	5	0.078*	0.153
19 9A - Nate's2	5	0.148*	0.199
20 9B - Nate's2	5	0.058*	0.214
21 10 - Drops 2	5	-0.203*	0.105
22 11 - Ken's 2	5	-0.040*	0.139

* Indicates that parameter is constrained.

[1] Indicates that difference between Grade 4 and Grade 5 parameters is statistically significant.

Table E.3

Parameter Estimates by Grade Level – Engaging in Argument from Evidence

Item	Grade	Estimate	Standard error
1 1A - Ana's ..	--	-0.022	0.122
2 1B - Ana's ..	--	-0.022	0.108
3 2A - Beth's..	--	0.133	0.109
4 2B - Beth's..	--	0.092	0.112
5 3A - Romita..	--	-0.163	0.096
6 3B - Romita..	--	-0.405	0.097
7 4A - Kevin'..	--	-0.096	0.118
8 4B - Kevin'..	--	-0.235	0.105
9 5A - Brass ..	--	0.010	0.119
10 5B - Brass ..	--	-0.220	0.103
11 6A - Carol'..	--	-0.056	0.115
12 6B - Carol'..	--	0.253	0.108
13 7A - Box of..	--	-0.168	0.126
14 7B - Box of..	--	-0.053	0.091
15 8A.1 - Grai..	--	-0.145	0.109
16 8A.2 - Grai..	--	0.000	0.132
17 8B.1 - Grai..	--	0.031	0.089
18 8B.2 - Grai..	--	0.173	0.092
19 9A - Nate's..	--	0.091	0.122
20 9B - Nate's..	--	-0.086	0.118
21 10 - Drops ..	--	0.274	0.080
22 11 - Ken's ..	--	0.272	0.083
--	4	0.022*	0.122
--	5	0.022*	0.108
1 1A - Ana's 1	4	-0.133*	0.109
2 1B - Ana's 1	4	-0.092*	0.112
3 2A - Beth's1	4	0.163*	0.096
4 2B - Beth's1	4	0.405*	0.097
5 3A - Romita1	4	0.096*	0.118
6 3B - Romita1	4	0.235*	0.105
7 4A - Kevin'1	4	-0.010*	0.119
8 4B - Kevin'1	4	0.220*	0.103
9 5A - Brass 1	4	0.056*	0.115
10 5B - Brass 1	4	-0.253*	0.108
11 6A - Carol'1	4	0.168*	0.126
12 6B - Carol'1	4	0.053*	0.091
13 7A - Box of1	4	0.145*	0.109

Item	Grade	Estimate	Standard error
14 7B - Box of1	4	-0.000*	0.132
15 8A.1 - Grai1	4	-0.031*	0.089
16 8A.2 - Grai1	4	-0.173*	0.092
17 8B.1 - Grai1	4	-0.091*	0.122
18 8B.2 - Grai1	4	0.086*	0.118
19 9A - Nate's1	4	-0.274*	0.080
20 9B - Nate's1	4	-0.272*	0.083
21 10 - Drops 1	4	-0.022	0.122
22 11 - Ken's 1	4	-0.022	0.108
1 1A - Ana's 2	5	0.133	0.109
2 1B - Ana's 2	5	0.092	0.112
3 2A - Beth's2	5	-0.163	0.096
4 2B - Beth's2	5	-0.405	0.097
5 3A - Romita2	5	-0.096	0.118
6 3B - Romita2	5	-0.235	0.105
7 4A - Kevin'2	5	0.010	0.119
8 4B - Kevin'2	5	-0.220	0.103
9 5A - Brass 2	5	-0.056	0.115
10 5B - Brass 2	5	0.253	0.108
11 6A - Carol'2	5	-0.168	0.126
12 6B - Carol'2	5	-0.053	0.091
13 7A - Box of2	5	-0.145	0.109
14 7B - Box of2	5	0.000	0.132
15 8A.1 - Grai2	5	0.031	0.089
16 8A.2 - Grai2	5	0.173	0.092
17 8B.1 - Grai2	5	0.091	0.122
18 8B.2 - Grai2	5	-0.086	0.118
19 9A - Nate's2	5	0.274	0.080
20 9B - Nate's2	5	0.272	0.083
21 10 - Drops 2	5	0.022*	0.122
22 11 - Ken's 2	5	0.022*	0.108

* Indicates that parameter is constrained

[1] Indicates that difference between Grade 4 and Grade 5 parameters is statistically significant.

Table E.4

Parameter Estimates by Inquiry Project Participation (5th Grade Students Only) – Structure and Properties of Matter

Item		Curriculum	Estimate	Standard error
1	1A - Ana's ..	--	-0.491	0.164
2	1B - Ana's ..	--	-1.221	0.244
3	2A - Beth's..	--	0.195	0.161
4	2B - Beth's..	--	0.216	0.157
5	3A - Romita..	--	-0.352	0.146
6	3B - Romita..	--	-0.239	0.141
7	4A - Kevin'..	--	0.497	0.141
8	4B - Kevin'..	--	0.097	0.128
9	5A - Brass ..	--	-0.303	0.145
10	5B - Brass ..	--	0.120	0.138
11	6A - Carol'..	--	0.397	0.153
12	6B - Carol'..	--	-0.197	0.145
13	7A - Box of..	--	1.755	0.232
14	7B - Box of..	--	0.663	0.182
15	8A - Grain ..	--	-0.055	0.161
16	8B - Grain ..	--	-1.101	0.227
17	9A - Nate's..	--	-0.589	0.155
18	9B - Nate's..	--	-0.826	0.181
19	10 - Drops ..	--	-0.465	0.096
20	11 - Ken's ..	--	-0.514	0.095
--		IP	-0.145[I]	0.033
--		non-IP	0.145*[I]	0.033
1	1A - Ana's 1	IP	0.433[I]	0.164
2	1B - Ana's 1	IP	0.711[I]	0.244
3	2A - Beth's1	IP	-0.158	0.161
4	2B - Beth's1	IP	-0.085	0.157
5	3A - Romita1	IP	-0.168	0.146
6	3B - Romita1	IP	-0.025	0.141
7	4A - Kevin'1	IP	0.153	0.141
8	4B - Kevin'1	IP	0.179	0.128
9	5A - Brass 1	IP	0.234	0.145
10	5B - Brass 1	IP	-0.209	0.138
11	6A - Carol'1	IP	0.006	0.153
12	6B - Carol'1	IP	-0.086	0.145
13	7A - Box of1	IP	-0.038	0.232
14	7B - Box of1	IP	0.299	0.182

Item	Grade	Estimate	Standard error
15 8A - Grain 1	IP	-0.574[I]	0.161
16 8B - Grain 1	IP	0.207	0.227
17 9A - Nate's1	IP	0.163	0.155
18 9B - Nate's1	IP	0.246	0.181
19 10 - Drops 1	IP	-0.460[I]	0.096
20 11 - Ken's 1	IP	0.234[I]	0.095
1 1A - Ana's 2	non-IP	-0.433*	0.164
2 1B - Ana's 2	non-IP	-0.711*	0.244
3 2A - Beth's2	non-IP	0.158*	0.161
4 2B - Beth's2	non-IP	0.085*	0.157
5 3A - Romita2	non-IP	0.168*	0.146
6 3B - Romita2	non-IP	0.025*	0.141
7 4A - Kevin'2	non-IP	-0.153*	0.141
8 4B - Kevin'2	non-IP	-0.179*	0.128
9 5A - Brass 2	non-IP	-0.234*	0.145
10 5B - Brass 2	non-IP	0.209*	0.138
11 6A - Carol'2	non-IP	-0.006*	0.153
12 6B - Carol'2	non-IP	0.086*	0.145
13 7A - Box of2	non-IP	0.038*	0.232
14 7B - Box of2	non-IP	-0.299*	0.182
15 8A - Grain 2	non-IP	0.574*	0.161
16 8B - Grain 2	non-IP	-0.207*	0.227
17 9A - Nate's2	non-IP	-0.163*	0.155
18 9B - Nate's2	non-IP	-0.246*	0.181
19 10 - Drops 2	non-IP	0.460*	0.096
20 11 - Ken's 2	non-IP	-0.234*	0.095

* Indicates that parameter is constrained

[I] Indicates that difference between *Inquiry Project* and non-*Inquiry* parameters is statistically significant.

Table E.5

Parameter Estimates by Inquiry Project Participation (5th Grade Students Only) – Scale, Proportion, and Quantity

Item		Curriculum	Estimate	Standard error
1	1A - Ana's ..	--	-1.491	0.292
2	1B - Ana's ..	--	-0.533	0.306
3	2A - Beth's..	--	-0.454	0.168
4	2B - Beth's..	--	-0.374	0.18
5	3A - Romita..	--	-0.689	0.262
6	3B - Romita..	--	-1.896	0.427
7	4A - Kevin'..	--	-2.021	0.334
8	4B - Kevin'..	--	-2.218	0.307
9	5A - Brass ..	--	-2.819	0.446
10	5B - Brass ..	--	-1.988	0.323
11	6A - Carol'..	--	-2.723	0.399
12	6B - Carol'..	--	-0.482	0.321
13	7A - Box of..	--	0.877	0.164
14	7B - Box of..	--	2.327	0.376
15	8A.1 - Grai..	--	0.596	0.224
16	8A.2 - Grai..	--	1.421	0.255
17	8B.1 - Grai..	--	-0.497	0.214
18	8B.2 - Grai..	--	0.252	0.186
19	9A - Nate's..	--	-0.151	0.259
20	9B - Nate's..	--	-0.472	0.291
21	10 - Drops ..	--	-1.872	0.145
22	11 - Ken's ..	--	-1.224	0.187
--		IP	-0.394[I]	0.050
--		non-IP	0.394*[I]	0.050
1	1A - Ana's 1	IP	0.234	0.292
2	1B - Ana's 1	IP	0.129	0.306
3	2A - Beth's1	IP	-0.016	0.168
4	2B - Beth's1	IP	0.184	0.180
5	3A - Romita1	IP	0.236	0.262
6	3B - Romita1	IP	0.824	0.427
7	4A - Kevin'1	IP	-0.399	0.334
8	4B - Kevin'1	IP	-0.790[I]	0.307
9	5A - Brass 1	IP	0.522	0.446
10	5B - Brass 1	IP	-0.164	0.323
11	6A - Carol'1	IP	-0.768	0.399
12	6B - Carol'1	IP	-0.210	0.321

Item	Grade	Estimate	Standard error
13 7A - Box of1	IP	0.021	0.164
14 7B - Box of1	IP	0.613	0.376
15 8A.1 - Grai1	IP	0.020	0.224
16 8A.2 - Grai1	IP	-0.061	0.255
17 8B.1 - Grai1	IP	-0.206	0.214
18 8B.2 - Grai1	IP	0.232	0.186
19 9A - Nate's1	IP	0.351	0.259
20 9B - Nate's1	IP	0.256	0.291
21 10 - Drops 1	IP	-0.373[ɪ]	0.145
22 11 - Ken's 1	IP	0.274	0.187
1 1A - Ana's 2	non-IP	-0.234*	0.292
2 1B - Ana's 2	non-IP	-0.129*	0.306
3 2A - Beth's2	non-IP	0.016*	0.168
4 2B - Beth's2	non-IP	-0.184*	0.180
5 3A - Romita2	non-IP	-0.236*	0.262
6 3B - Romita2	non-IP	-0.824*	0.427
7 4A - Kevin'2	non-IP	0.399*	0.334
8 4B - Kevin'2	non-IP	0.790*[ɪ]	0.307
9 5A - Brass 2	non-IP	-0.522*	0.446
10 5B - Brass 2	non-IP	0.164*	0.323
11 6A - Carol'2	non-IP	0.768*	0.399
12 6B - Carol'2	non-IP	0.210*	0.321
13 7A - Box of2	non-IP	-0.021*	0.164
14 7B - Box of2	non-IP	-0.613*	0.376
15 8A.1 - Grai2	non-IP	-0.020*	0.224
16 8A.2 - Grai2	non-IP	0.061*	0.255
17 8B.1 - Grai2	non-IP	0.206*	0.214
18 8B.2 - Grai2	non-IP	-0.232*	0.186
19 9A - Nate's2	non-IP	-0.351*	0.259
20 9B - Nate's2	non-IP	-0.256*	0.291
21 10 - Drops 2	non-IP	0.373*[ɪ]	0.145
22 11 - Ken's 2	non-IP	-0.274*	0.187

* Indicates that parameter is constrained

[ɪ] Indicates that difference between *Inquiry Project* and non-*Inquiry* parameters is statistically significant.

Table E.6

Parameter Estimates by Inquiry Project Participation (5th Grade Students Only) – Engaging in Argument from Evidence

Item		Curriculum	Estimate	Standard error
1	1A - Ana's ..	--	-1.150	0.168
2	1B - Ana's ..	--	-0.583	0.145
3	2A - Beth's..	--	-1.247	0.142
4	2B - Beth's..	--	-1.338	0.160
5	3A - Romita..	--	-0.764	0.121
6	3B - Romita..	--	-0.657	0.125
7	4A - Kevin'..	--	-1.003	0.152
8	4B - Kevin'..	--	-1.049	0.141
9	5A - Brass ..	--	-0.946	0.169
10	5B - Brass ..	--	-0.800	0.135
11	6A - Carol'..	--	-1.511	0.147
12	6B - Carol'..	--	-1.068	0.139
13	7A - Box of..	--	-0.772	0.170
14	7B - Box of..	--	-0.680	0.113
15	8A.1 - Grai..	--	-0.954	0.143
16	8A.2 - Grai..	--	-0.699	0.162
17	8B.1 - Grai..	--	-0.630	0.118
18	8B.2 - Grai..	--	-0.210	0.110
19	9A - Nate's..	--	-0.150	0.170
20	9B - Nate's..	--	-1.117	0.167
21	10 - Drops ..	--	-1.164	0.112
22	11 - Ken's ..	--	-1.427	0.114
--		IP	-0.165[1]	0.029
--		non-IP	0.165*[1]	0.029
1	1A - Ana's 1	IP	0.106	0.168
2	1B - Ana's 1	IP	-0.246	0.145
3	2A - Beth's1	IP	-0.006	0.142
4	2B - Beth's1	IP	-0.310	0.160
5	3A - Romita1	IP	-0.026	0.121
6	3B - Romita1	IP	-0.427[1]	0.125
7	4A - Kevin'1	IP	-0.101	0.152
8	4B - Kevin'1	IP	0.214	0.141
9	5A - Brass 1	IP	0.058	0.169
10	5B - Brass 1	IP	-0.005	0.135
11	6A - Carol'1	IP	-0.075	0.147
12	6B - Carol'1	IP	0.355[1]	0.139

Item	Grade	Estimate	Standard error
13 7A - Box of1	IP	-0.082	0.170
14 7B - Box of1	IP	-0.089	0.113
15 8A.1 - Grai1	IP	-0.053	0.143
16 8A.2 - Grai1	IP	0.124	0.162
17 8B.1 - Grai1	IP	-0.354[ɪ]	0.118
18 8B.2 - Grai1	IP	-0.160	0.110
19 9A - Nate's1	IP	1.011[ɪ]	0.170
20 9B - Nate's1	IP	0.097	0.167
21 10 - Drops 1	IP	0.089	0.112
22 11 - Ken's 1	IP	0.198	0.114
1 1A - Ana's 2	non-IP	-0.106*	0.168
2 1B - Ana's 2	non-IP	0.246*	0.145
3 2A - Beth's2	non-IP	0.006*	0.142
4 2B - Beth's2	non-IP	0.310*	0.160
5 3A - Romita2	non-IP	0.026*	0.121
6 3B - Romita2	non-IP	0.427*[ɪ]	0.125
7 4A - Kevin'2	non-IP	0.101*	0.152
8 4B - Kevin'2	non-IP	-0.214*	0.141
9 5A - Brass 2	non-IP	-0.058*	0.169
10 5B - Brass 2	non-IP	0.005*	0.135
11 6A - Carol'2	non-IP	0.075*	0.147
12 6B - Carol'2	non-IP	-0.355*[ɪ]	0.139
13 7A - Box of2	non-IP	0.082*	0.170
14 7B - Box of2	non-IP	0.089*	0.113
15 8A.1 - Grai2	non-IP	0.053*	0.143
16 8A.2 - Grai2	non-IP	-0.124*	0.162
17 8B.1 - Grai2	non-IP	0.354*[ɪ]	0.118
18 8B.2 - Grai2	non-IP	0.160*	0.110
19 9A - Nate's2	non-IP	-1.011*[ɪ]	0.170
20 9B - Nate's2	non-IP	-0.097*	0.167
21 10 - Drops 2	non-IP	-0.089*	0.112
22 11 - Ken's 2	non-IP	-0.198*	0.114

* Indicates that parameter is constrained

[ɪ] Indicates that difference between Grade 4 and Grade 5 parameters is statistically significant.

Appendix F

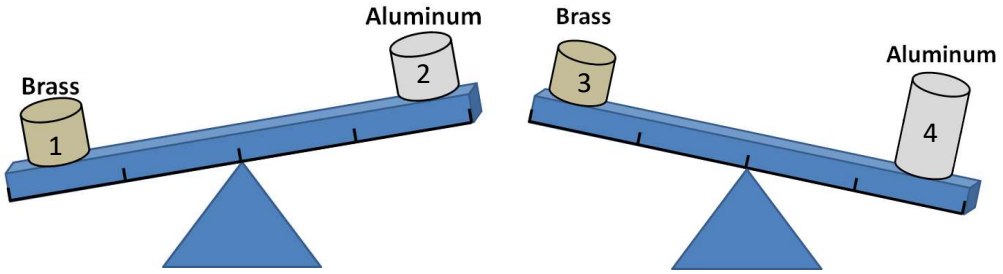
Pilot Assessment Items

Brass and Aluminum A – Constructed Response Argument

There are four cylinders on two balance scales. They are made of two different materials. Two cylinders are made of a material called brass. Two cylinders are made of a material called aluminum.

On Scale 1, cylinders 1 and 2 have the same volume.

On Scale 2, cylinders 3 and 4 have different volumes.



Volume: The amount of space that something takes up

1a) What can you tell about the weight of cylinders 1 and 2?

- ☐ Cylinder 1 weighs more.
- ☐ Cylinder 2 weighs more.
- ☐ The cylinders weigh the same amount.
- ☐ You cannot tell which one weighs more.

1b) What can you tell about the weight of cylinders 3 and 4?

- ☐ Cylinder 3 weighs more.
- ☐ Cylinder 4 weighs more.
- ☐ The cylinders weigh the same amount.
- ☐ You cannot tell which one weighs more.

A very small piece of brass is the exact same size and shape as another very small piece of aluminum.

2a) What can you tell about the weight of these small pieces of brass and aluminum?

- ☐ The brass piece will weigh more.
- ☐ The aluminum piece will weigh more.
- ☐ They will both weigh the same tiny bit.
- ☐ They will both weigh nothing at all.
- ☐ You cannot tell anything about their weight.

2b) How can you tell? Make an argument. Give your evidence and reasoning.

_____

_____

_____

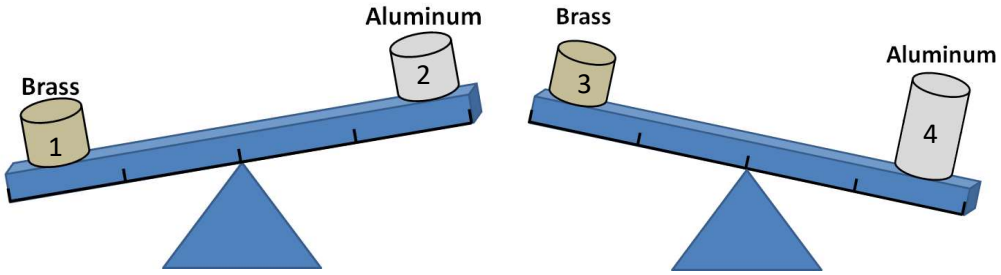
_____

Brass and Aluminum B – Selected Response Argument

There are four cylinders on two balance scales. They are made of two different materials. Two cylinders are made of a material called brass. Two cylinders are made of a material called aluminum.

On Scale 1, cylinders 1 and 2 have the same volume.

On Scale 2, cylinders 3 and 4 have different volumes.



Volume: The amount of space that something takes up

1a) What can you tell about the weight of cylinders 1 and 2?

- ☐ Cylinder 1 weighs more.
- ☐ Cylinder 2 weighs more.
- ☐ The cylinders weigh the same amount.
- ☐ You cannot tell which one weighs more.

1b) What can you tell about the weight of cylinders 3 and 4?

- ☐ Cylinder 3 weighs more.
- ☐ Cylinder 4 weighs more.
- ☐ The cylinders weigh the same amount.
- ☐ You cannot tell which one weighs more.

A very small piece of brass is the exact same size and shape as another very small piece of aluminum.

Brass **Aluminum**

2a) What can you tell about the weight of these small pieces of brass and aluminum?

- ☐ The brass piece will weigh more.
- ☐ The aluminum piece will weigh more.
- ☐ They will both weigh the same tiny bit.
- ☐ They will both weigh nothing at all.
- ☐ You cannot tell anything about their weight.

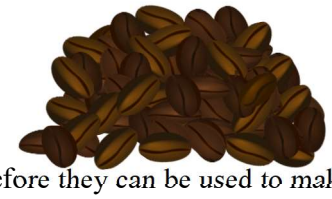
2b) How can you tell? Make an argument. Give your evidence and reasoning. You can pick more than one answer.

- ☐ I can tell because the new pieces are the same size and shape.
- ☐ I can tell because the new pieces are both very small.
- ☐ I can tell because of the weight of equal volumes of brass and aluminum on Scale 1.
- ☐ I can tell because of the weight of different volumes of brass and aluminum on Scale 2.
- ☐ I cannot tell because the pieces are different sizes than the pieces on Scale 1 and 2.
- ☐ Very small things will weigh only a tiny bit.
- ☐ Very small things do not weigh anything.
- ☐ Objects that are the same size should always weigh the same.
- ☐ A piece made of brass will always be heavier than a piece made of aluminum when the two pieces are the same size and shape.
- ☐ A piece made of aluminum will always be heavier than a piece made of brass when the two pieces are the same size and shape.
- ☐ I do not think there is enough evidence. The evidence I need is: _____
- ☐ I had some other reason: _____

Measuring coffee A – Constructed-response (all prompts)

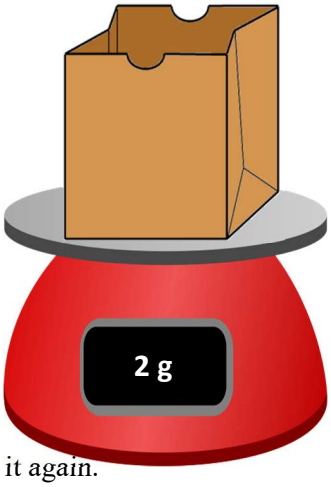
Coffee starts out in the form of beans.

a coffee bean

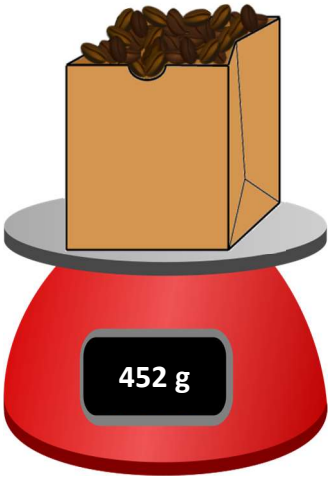


The beans must be crushed into powder before they can be used to make drinks. This is what coffee looks like after the beans have been crushed into powder.

Kevin has an empty paper bag. He weighs it on a scale.



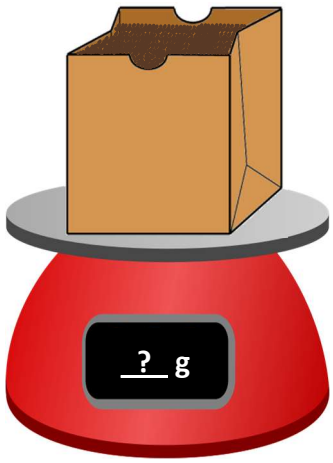
Then he fills the bag with beans and weighs it again.



1) How much do *just* the beans weigh?

_____ grams

Kevin uses a machine to crush the beans. He is very careful not to lose any pieces of the powder.



2a) How much do you think the bag of powder weighs?

_____ grams

2b) Why do you think so? Make an argument. Give your evidence and reasoning.

_____

_____

_____

_____

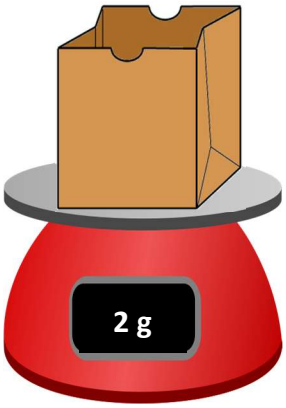
Measuring coffee B – Selected-response (all prompts)

Coffee starts out in the form of beans.

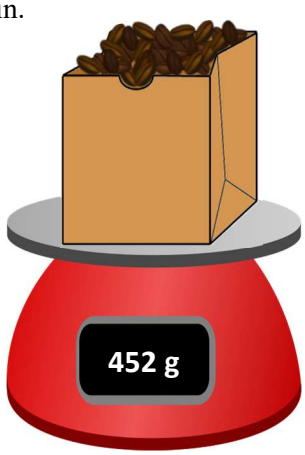

a coffee bean

The beans must be crushed into powder before they can be used to make drinks. This is what coffee looks like after the beans have been crushed into powder.

Kevin has an empty paper bag. He weighs it on a scale.



Then he fills the bag with beans and weighs it again.



1) How much do *just* the beans weigh?

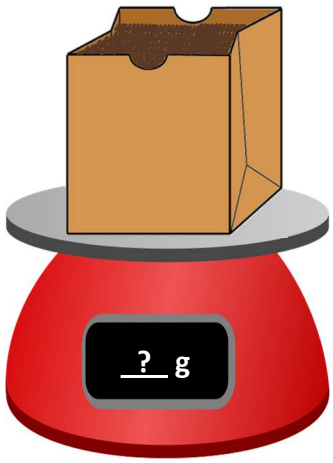
☐ 450 grams

☐ 452 grams

☐ 454 grams

☐ You cannot tell how much they weigh

Kevin uses a machine to crush the beans. He is very careful not to lose any pieces of the powder.



2a) How much do you think the bag of powder weighs?

- ☐ Much less than 452 grams
- ☐ A little less than 452 grams
- ☐ 452 grams
- ☐ A little more than 452 grams
- ☐ Much more than 452 grams

2b) Why do you think so? Make an argument. Give your evidence and reasoning. You can choose more than one answer.

- ☐ The coffee changed its form from beans to powder.
- ☐ Kevin didn't add or lose any coffee.
- ☐ The powder will take up less space in the bag than the beans.
- ☐ The coffee changed form, so the weight will change.
- ☐ The amount of coffee stayed the same, so the weight will stay the same.
- ☐ The amount of space the coffee takes up changed, so the weight will change.
- ☐ I do not think there is enough evidence. The evidence I need is: _____
- ☐ I had some other reason: _____

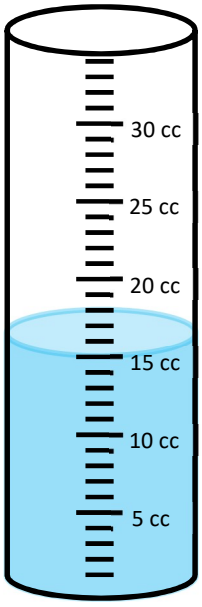
Beth's rock A – Constructed-response argument

Beth has a rock.

She has a measuring cup. She pours some water into the measuring cup. Then she puts the rock in the measuring cup of water.

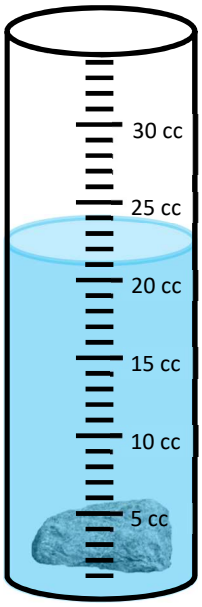
Picture A shows what the measuring cup looks like before she puts the rock into it.

Picture A



Picture B shows what happens after she puts the rock into the measuring cup.

Picture B



1a) From the pictures, can you tell what the volume of the rock is?

Volume: The amount of space that something takes up

☐ Yes, the volume of the rock is _____.

☐ No. What other information do you need? _____

1b) Why or why not? Make an argument. Give your evidence and reasoning.

_____

_____

_____

_____

2a) From the pictures, can you tell what the weight of the rock is?

☐ Yes, the weight of the rock is _____.

☐ No. What other information do you need? _____

2b) Why or why not? Make an argument. Give your evidence and reasoning.

_____

_____

_____

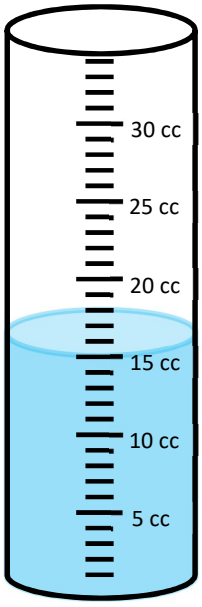
_____

Beth's rock B – Selected-response argument

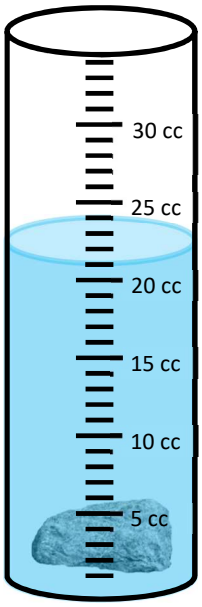
Beth has a rock.

She has a measuring cup. She pours some water into the measuring cup. Then she puts the rock in the measuring cup of water.

Picture A shows what the measuring cup looks like before she puts the rock into it.



Picture B shows what happens after she puts the rock into the measuring cup.



1a) From the pictures, can you tell what the volume of the rock is?

Volume: The amount of space that something takes up

☐ Yes, the volume of the rock is _____.

☐ No. What other information do you need? _____

1b) Why or why not? Make an argument. Give your evidence and reasoning. You can pick more than one answer.

- ☐ I can tell because the water level went up.
- ☐ I can tell because of the way the rock looks.
- ☐ The amount that the water level rises depends on the rock's volume.
- ☐ The amount that the water level rises depends on the rock's weight.
- ☐ You can tell how much space something takes up by looking at it.
- ☐ I don't think there is any evidence of the rock's volume in the picture.
  The evidence I need is:
  _____
- ☐ I had some other reason: _____

2a) From the pictures, can you tell what the weight of the rock is?

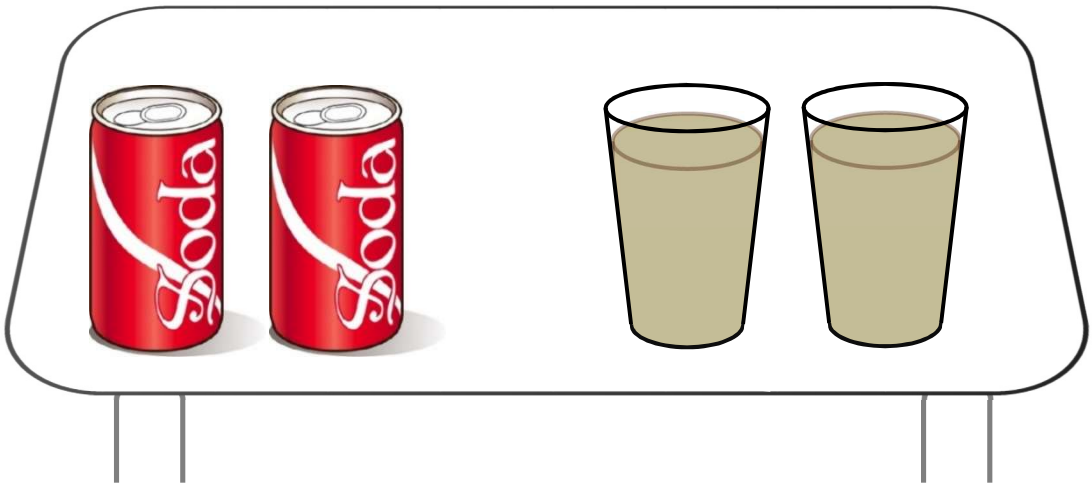
- ☐ Yes, the weight of the rock is _____.
- ☐ No. What other information do you need? _____

2b) Why or why not? Make an argument. Give your evidence and reasoning. You can pick more than one answer.

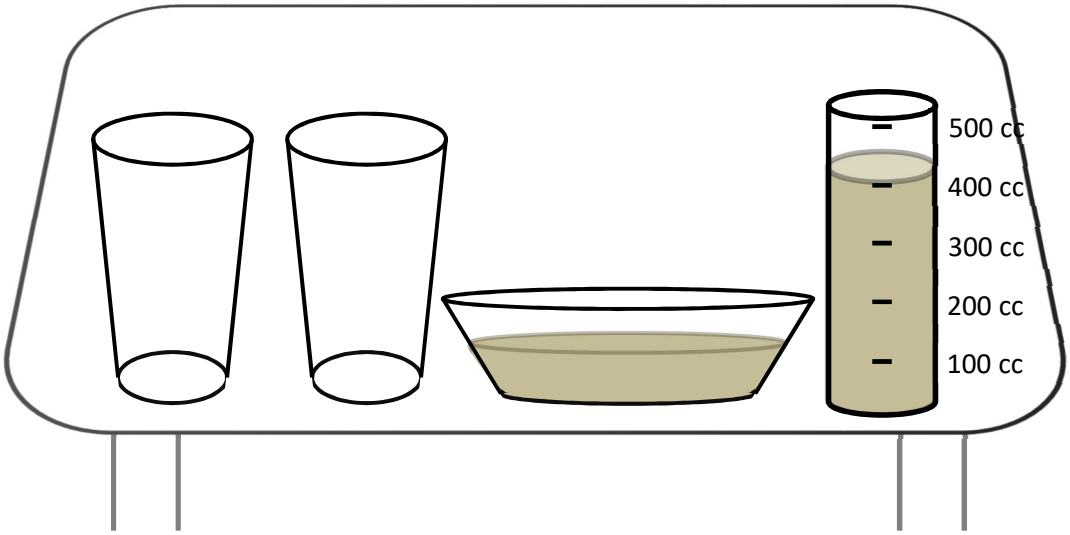
- ☐ I can tell because of the amount that the water level rose.
- ☐ I can tell because of the way the rock looks.
- ☐ The amount that the water level rises depends on the rock's weight.
- ☐ The amount that the water level rises depends on the rock's volume.
- ☐ You can tell how much space something takes up by looking at it.
- ☐ I don't think there is any evidence of the rock's volume in the picture.
  The evidence I need is:
  _____
- ☐ I had some other reason: _____

Cans of soda A – Constructed-response argument

Nate pours two cans of soda into two glasses.



He pours all of the soda from one glass into a shallow bowl. He pours the other glass into a tall measuring cup.



1a) Is the volume of soda in the bowl and the measuring cup the same, or different?

☐ Same.

☐ Different.

Volume:
The amount of space that something takes up

1b) Why do you think so? Make an argument. Give your evidence and reasoning.

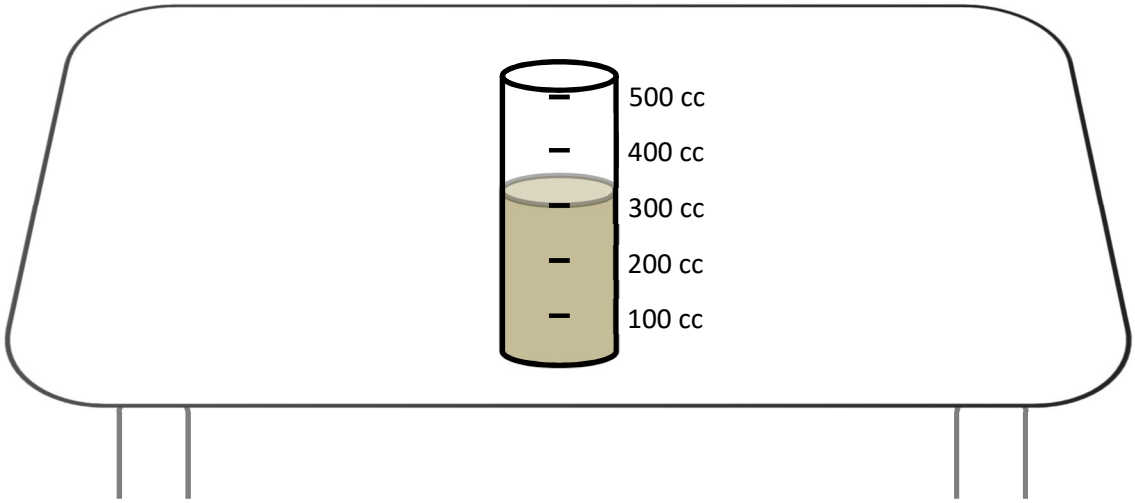
_____

_____

_____

_____

Nate pours some soda out of the measuring cup.



2) Does the volume of soda in the measuring cup change?

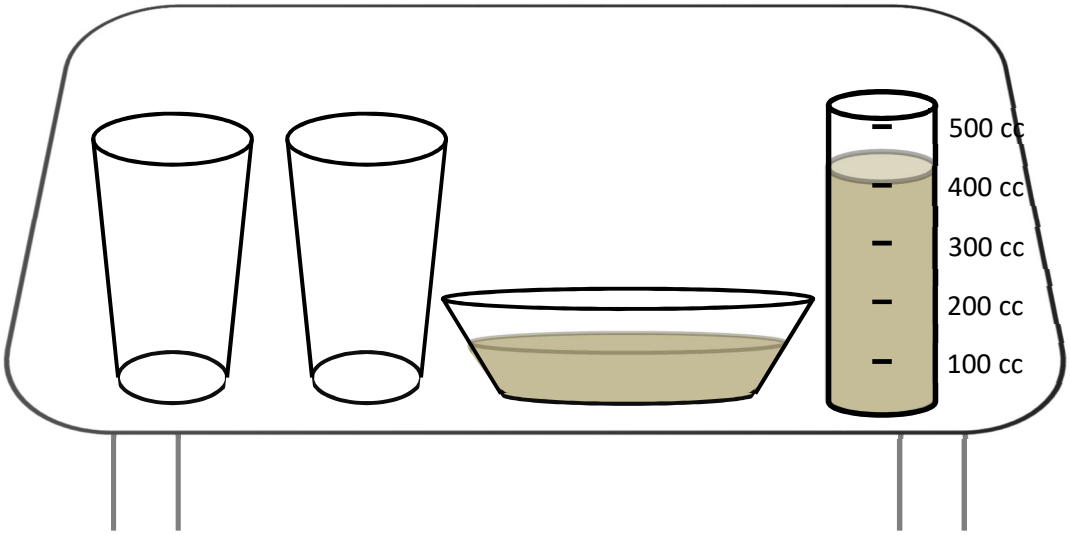
☐ Yes. What is the volume of the soda that Nate poured out? _____

☐ No. Why not? _____

Cans of soda B – Selected-response argument

Nate pours two cans of soda into two glasses.

He pours all of the soda from one glass into a shallow bowl. He pours the other glass into a tall measuring cup.



1a) Is the volume of soda in the bowl and the measuring cup the same, or different?

☐ Same.

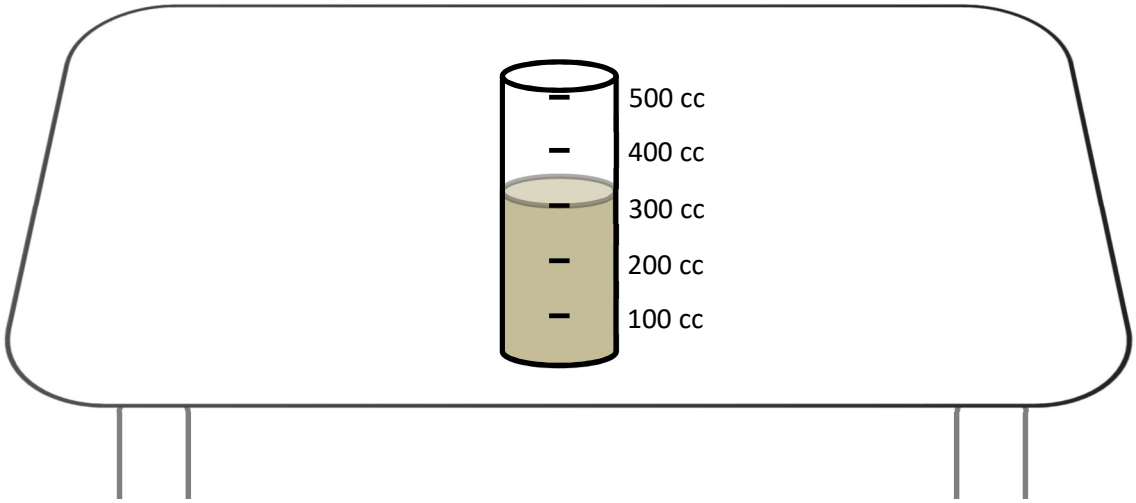
☐ Different.

Volume:
The amount of space that something takes up

1b) Why do you think so? Make an argument. Give your evidence and reasoning. You can pick more than one answer.

- ☐ The containers are different.
- ☐ The height of the soda is different.
- ☐ Nate poured the same amount of soda into each container.
- ☐ The soda takes up a different amount of space in the containers.
- ☐ The containers take up different amounts of space on the table.
- ☐ The same amount of soda will take up the same amount of space.
- ☐ I had some other reason: _____

Nate pours some soda out of the measuring cup and into the sink.



2) Does the volume of soda in the measuring cup change?

- ☐ Yes. How much did it change?
  _____
- ☐ No. Why not? _____

Carol's butter A – Multiple-prompt

Carol puts a chunk of cold butter into a small bowl. The empty bowl weighs 5 grams.

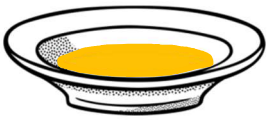
The bowl with the cold butter weighs 7 grams.

1) How much does the chunk of cold butter weigh?

_____ grams

She heats the cold butter until it melts. The butter becomes liquid. None of the butter burns or evaporates.

2a) How much do you think the melted butter weighs?



_____ grams

2b) How do you know? Make an argument. Give your evidence and reasoning.

_____

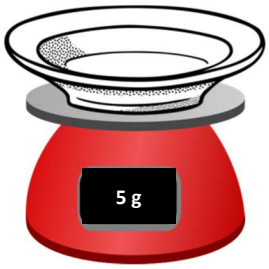
_____

_____

_____

Carol's butter B – Single-prompt

Carol puts a chunk of cold butter into a small bowl. The empty bowl weighs 5 grams.



The bowl with the cold butter weighs 7 grams.

She heats the cold butter until it melts. The butter becomes liquid. None of the butter burns or evaporates.

How much do you think *just* the melted butter weighs? How do you know? Make an argument. Give your evidence and reasoning.

_____

_____

_____

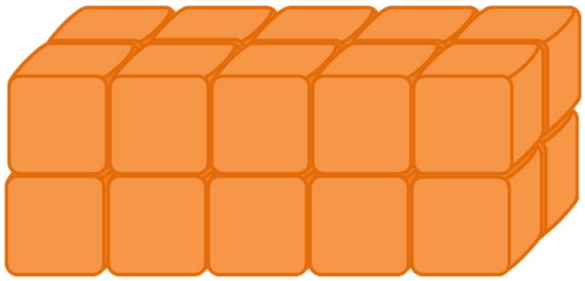
_____

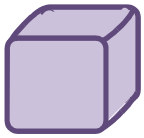
_____

_____

Ana's block of clay A – Multiple-prompt

Ana has a block of clay. The block of clay is marked so that it can be divided into smaller pieces. Each smaller piece is 1 cubic centimeter.





1 cubic centimeter

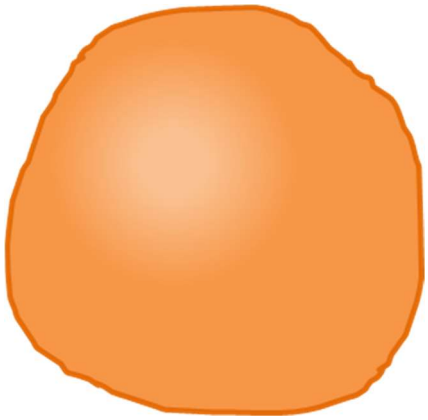
1) What i

block of clay?

Volume:
The amount of space that something takes up

_____

Ana takes the block of clay and molds it into a ball. She is careful not to get any air inside of the ball of clay.



2a) What is the volume of the ball of clay?

_____

2b) Why do you think so? Make an argument. Give your evidence and reasoning.

_____

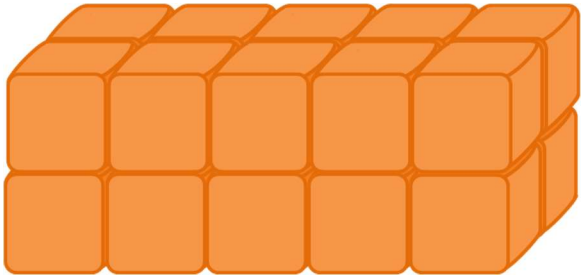
_____

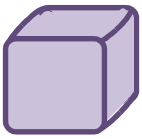
_____

_____

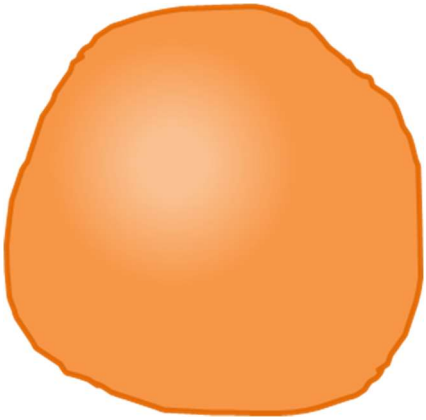
Ana's block of clay B – Single-prompt

Ana has a block of clay. The block of clay is marked so that it can be divided into smaller pieces. Each smaller piece is 1 cubic centimeter.



1 cubic centimeter 

Ana takes the block of clay and molds it into a ball. She is careful not to get any air inside of the ball of clay.

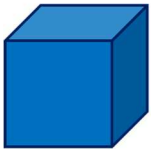
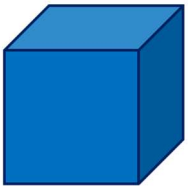
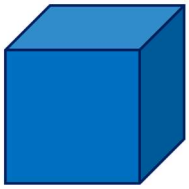
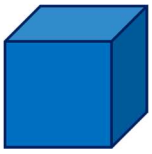


1) What is the volume of the ball of clay? Why do you think so? Make an argument. Give your evidence and reasoning.

Volume:
The amount of space that something takes up

Four blocks A – Multiple-prompt

Here are four blocks. All of the blocks are solid, with no air inside of them. The chart shows the weight and volume of each block.

	A	B	C	D
				
Weight (grams)	5 g	15 g	10 g	10 g
Volume (cubic centimeters)	5 cc	10 cc	10 cc	5 cc

1) Which block takes up the most space? (You can pick more than one block if there is a tie.)

_____

Volume: The amount of space that something takes up

2) Which block is heaviest? (You can pick more than one block if there is a tie.)

_____

3) Which blocks are heaviest for their sizes?

_____

4a) Do you think any of the blocks could be made of the same material?

☐ Yes. Which ones? _____

☐ No, none of the blocks could be made of the same material.

4b) Why do you think so? Make an argument. Give your evidence and reasoning.

_____

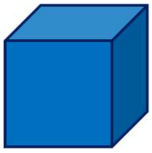
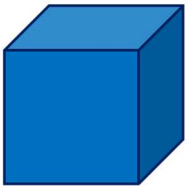
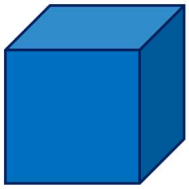
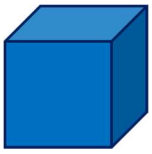
_____

_____

_____

Four blocks B – Single-prompt

Here are four blocks. All of the blocks are solid, with no air inside of them. The chart shows the weight and volume of each block.

	A	B	C	D
				
Weight (grams)	5 g	15 g	10 g	10 g
Volume (cubic centimeters)	5 cc	10 cc	10 cc	5 cc

1) Do you think any of the blocks could be made of the same material? Why or why not? Make an argument. Give your evidence and reasoning.

Volume: The amount of space that something takes up

Material: Glass, sand, and different types of wood and metal are examples of different materials

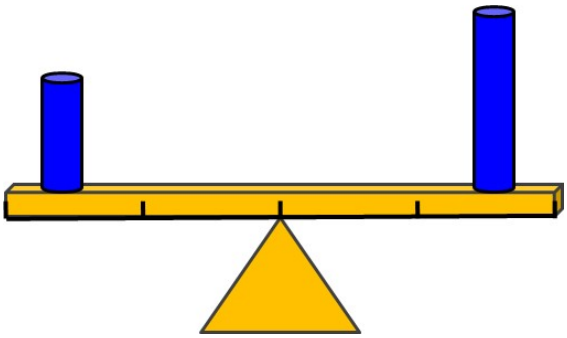
Romita's cylinders A – Multiple-prompt

Romita holds two cylinders. They are both completely solid. The tall cylinder has a larger volume than the short cylinder.

Volume:
The amount of space that something takes up

She says "The short cylinder feels heavier than the tall cylinder."

Romita puts both cylinders onto a scale. The scale balances perfectly.



1) Can you tell if one of the cylinders is heavier than the other?

- ☐ The tall cylinder is heavier.
- ☐ The short cylinder is heavier.
- ☐ They weigh the same.
- ☐ You cannot tell which cylinder is heavier.

2a) Could the cylinders be made of the same kind of material?

☐ Yes

☐ No

Material: Glass, sand, and different types of wood and metal are examples of different materials

2b) How can you tell? Make an argument. Give your evidence and reasoning.

_____

_____

_____

_____

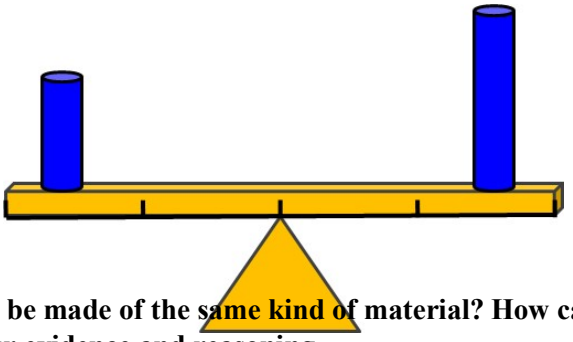
Romita's cylinders B – Single-prompt

Romita holds two cylinders. They are both completely solid. The tall cylinder has a larger volume than the short cylinder.

Volume:
The amount of space that something takes up

She says "The short cylinder feels heavier than the tall cylinder."

Romita puts both cylinders onto a scale. The scale balances perfectly.



1) Could the cylinders be made of the same kind of material? How can you tell? Make an argument. Give your evidence and reasoning.

Material: Glass, sand, and different types of wood and metal are examples of different materials

_____

_____

_____

_____

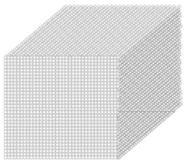
_____

_____

_____

Grain of sugar A – Multiple-prompt

1 cubic centimeter of sugar weighs 2 grams and contains 2000 grains of sugar.



1a) Does 1 grain of sugar weigh anything?

☐ Yes. How much does a grain of sugar weigh? _____

☐ No. About many grains of sugar would you need for it to weigh something? _____

1b) Why do you think so? Make an argument. Give your evidence and reasoning.

_____

_____

_____

_____

2a) Does 1 grain of sugar take up any space?

Volume:
The amount of space that something takes up

☐ Yes. What is the volume of a grain of sugar? _____

☐ No. About how many grains of sugar would you need for it to take up space? _____

2b) Why do you think so? Make an argument. Give your evidence and reasoning.

_____

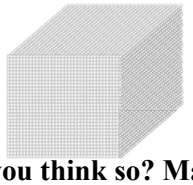
_____

_____

_____

Grain of sugar B – Single-prompt

1 cubic centimeter of sugar weighs 2 grams and contains 2000 grains of sugar.



How much does 1 grain of sugar weigh? Why do you think so? Make an argument. Give your evidence and reasoning.

_____

_____

_____

_____

_____

_____

_____

What is the volume of a grain of sugar? Why do you think so? Make an argument. Give your evidence and reasoning.

_____

_____

_____

Volume:
The amount of space that something takes up

_____

_____

_____

_____

_____

Water drops (This is the only variation given to students)

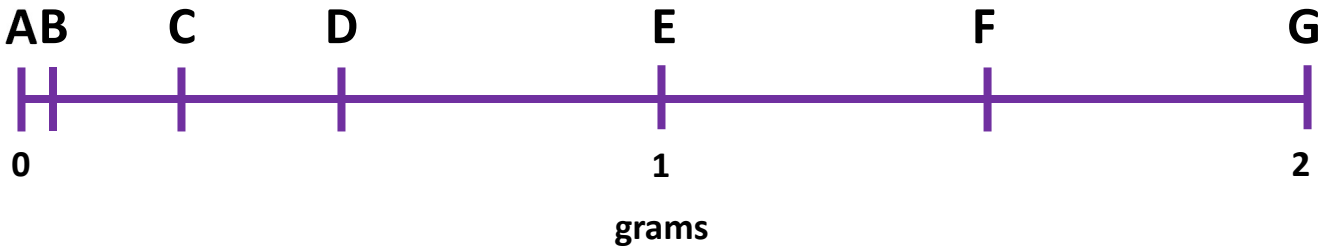
20 drops of water weigh 1 gram.

1a) Will 10 drops of water weigh anything?

☐ Yes.

☐ No.

1b) Circle the letter that marks the weight of 10 drops of water on the number line.

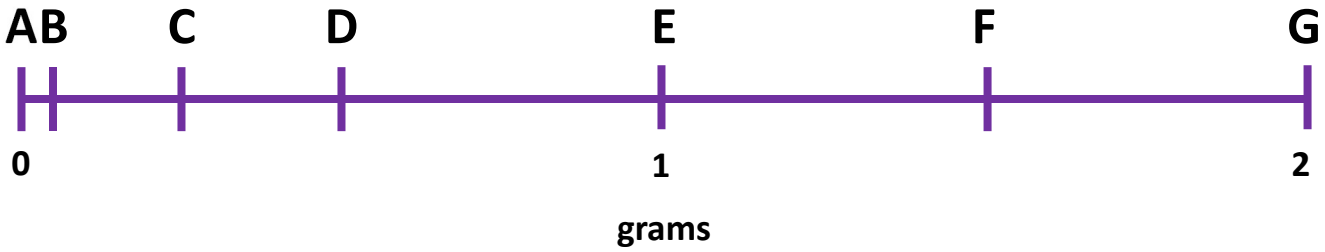


2a) Does 1 drop of water weigh anything?

☐ Yes

☐ No

2b) Circle the letter that marks the weight of 1 drop of water on the number line.



3) Why did you choose your answers? Make an argument. Give your evidence and reasoning.

_____

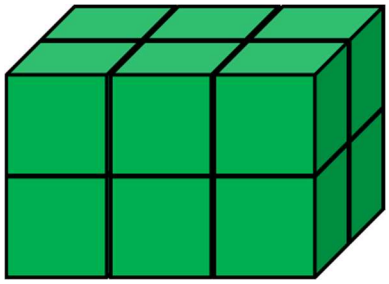
_____

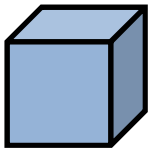
_____

_____

Block made of small cubes (This is the only variation given to students)

Ken uses some small cubes to make a big block. Each small cube is 1 cubic centimeter.





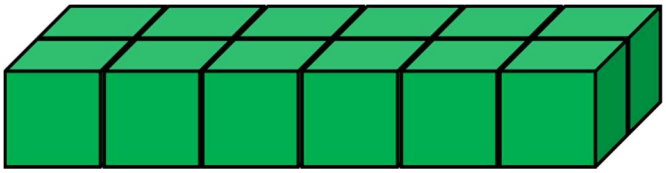
1 cubic centimeter

What is the volume of the big block?

_____

Volume:
The amount of space that something takes up

Ken moves the small cubes around to make a longer block.



Is the volume of the longer block the same as the first block, or different?

☐ Same

☐ Different

Why do you think so? Give your evidence and reasoning.

_____

_____

_____

_____