Boston College
Lynch School of Education and Human Development

Department of
Measurement, Evaluation, Statistics, and Assessment

MEASURING STUDENTS' PERCEPTIONS OF STUDENT TEACHING UNIVERSITY SUPERVISORS: SCENARIO-BASED SCALE DEVELOPMENT USING RASCH AND GUTTMAN FACET THEORY

Dissertation
by

KEVIN RICHARD HOLBROOK

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

May 2022

ABSTRACT

MEASURING STUDENTS' PERCEPTIONS OF STUDENT TEACHING UNIVERSITY SUPERVISORS: SCENARIO-BASED SCALE DEVELOPMENT USING RASCH AND GUTTMAN FACET THEORY

Kevin R. Holbrook, Author

Larry H. Ludlow, Chair

In the field of teacher education, it is well documented that the most influential part is the clinical component, often referred to as student teaching or the practicum experience (Cochran-Smith, 1991; Darling-Hammond, 2014; Evertson, 1990). During the practicum, there exists a triad of three individuals: a university supervisor, a K-12 classroom teacher (often referred to as the *cooperating teacher*), and the teacher candidate. While much research has been conducted on teacher candidates and the cooperating teacher role, there has been a lack of research on the role of the university supervisor.

The lack of measurement instruments to assess the quality of the university supervisor puts teacher education programs at a disadvantage, both from a programmatic improvement standpoint, as well as meeting accreditation requirements. This dissertation provides an answer to this issue, creating a new instrument that assesses the quality of the university supervisor, from the perspective of teacher candidates. This instrument is constructed under the Rasch-Guttman Scenario (RGS) framework, as most clearly defined by Ludlow, Baez-Cruz, et al. (2020). The RGS framework derives its influence from the works of George Rasch (1960/80) as well as Louis Guttman's facet theory (Guttman 1954; Guttman 1957). The result is a new scale, entitled the University Supervisor Quality (USQ) scale, consisting of nine scenario items. All nine items include

four facets that comprise the construct of university supervisor quality: resourcefulness, constructive feedback, mentorship, and collaboration.

The results of this dissertation suggest that the utilization of the RGS framework is successful for developing a scale about university supervisor quality. In addition, the use of cognitive interviews provide valuable insight into the development of scales using the RGS framework. This scale has the potential for use in teacher education programs for evaluating the quality of their supervisors, and to utilize as evidence for accreditation purposes.

DEDICATION

Mom, Dad, & Kerri.

I love you.

This is for you.

ACKNOWLEDGEMENTS

There are so many individuals I would like to thank for being a part of this dissertation. I entered Boston College in Fall 2011 as an undergraduate student majoring in history, and have somehow ended up here in 2022, finishing my Ph.D. in Measurement, Evaluation, Statistics, and Assessment. It has been an incredible journey.

I first want to thank my dissertation chair, Dr. Larry Ludlow. You have been an incredible teacher and advisor to me over these last few years. From our time working together on the CÖC accreditation team, conference presentations, classes, and ultimately this dissertation, I have learned so much from you (and your admitted "better half," Dr. Marilyn Cochran-Smith). Your influence on me as a researcher is without question.

I would also like to thank my other dissertation committee members. Dr. Nathaniel Brown, it was your class I took as a senior at BC (Classroom Assessment) that excited me about educational research, and made me want to continue with the MESA program. Dr. Amy Ryan, I want to thank you for all your help and support, both as a research assistant in your office and on my committee. Lastly, to Dr. Christine Power, I am so incredibly thankful for the many years of collaboration we have had, both in the Practicum Office and on my dissertation. The Medfield Public Schools are so incredibly lucky to have you supporting its teachers—I cannot think of a better person for that role.

This dissertation would not be possible without the wonderful people of the Office of Field Placement and Partnership Outreach, formerly the "Practicum Office." It was back in 2015 that I was hired by the lovely Fran Loftus, who encouraged me and inspired me to make the leap from the master's to the doctoral program. Thank you, Fran for taking a chance on me.

I want to thank my assistantship advisor over the last two years, Dr. Melodie Wyttenbach at the Roche Center for Catholic Education. You have been an incredible mentor, and I thoroughly enjoyed our work together. Thank you as well to the rest of my Roche Center research partners, Dr. Andrew Miller, Dr. John Reyes, Dr. Michael O'Connor, and Dr. Hoffsman Ospino. You have all taught me so much about the field of Catholic education.

While not a part of my academic career at BC, I also want to thank my RD of 3 years, Matt Burke, when I was an RA in the Mods/Stayer community. I loved being an RA, and Matt was the driving force in making that possible. And a special thank you to Matt's boss, Dorrie Siqueiros, who randomly assigned two RAs to work together for an assignment at Duchene Hall on Newton Campus.

There are also so many friends I am beyond grateful for. To my best friends from my undergraduate BC days, the "Washed Up Eagles." Thank you, Dan, Lucas, Ben, Chris, and Kevin, as well as Kristina and Annie, for always being that rock of support. And to my friends from Medfield, BC, and beyond, I cannot thank you enough for your friendship.

To my three greatest friends, Paul, Kabir, and Kyle. I am so incredibly proud of all three of you. I am so glad that while we all live in different parts across this nation, not a day goes by without the four of us talking. I love you guys.

I also want to thank the wonderful Keeler family—Joe, MaryAnn, Melissa, Brittany, and Alex (and Baloo!). You all have welcomed me with open arms and I feel very lucky to have you in my life. I truly love you all.

And while friends and colleagues surely are important, as Nana Holbrook used to say, “Family comes first.” Thank you to the extended Holbrook family, especially to those whose influence as educators and researchers have helped me on this dissertation. To Aunt Gail, Aunt Cathy (the original Dr. Holbrook), Uncle Paul, Uncle Jim, and my cousins MaryKate, Ryan, and Gina. Thank you for your thoughtful perspectives, insights, and support all these years.

I would like to thank my two grandmothers, Nana Wilson, and Nana Holbrook. Not every grandchild can say their grandmas were best friends—but I can. I lost my Nana Holbrook during this dissertation, and I miss her lovely Irish wit and wisdom every day. I am glad that you, Nana Wilson, are here to see me finish. I hope that you both, as well as Pops and PopPop are proud of me.

Finally, I want to thank the three most important people in my life: my parents Susan and Rich, and my incredible partner in life, Kerri. You all have given me unwavering support all these years, and quite frankly, this project would not have finished without you. But beyond this dissertation…

Mom and Dad, you have always believed in me, without question. You have been the greatest example of what it means to be generous, loving people. I am so proud to call you not just my parents, but my heroes. I love you both.

But you, Kerri.

First, I think we should thank Dorrie for pairing us together for our RA assignment at “Due-chess-knee” Hall on Newton Campus. Because just like Renée Zellweger said in *Jerry Maguire*, you really did “have me at hello.” You are, my person. I love you.

And our Sully-Boy!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE: INTRODUCTION

The Problem

The field of teacher preparation has experienced tremendous public scrutiny and political attention during the last three decades. The National Commission on Excellence in Education's release of *A Nation at Risk* (1983) was a stinging report on the state of American public education, which propelled forth a long-lasting narrative of an inadequate public educational system riddled with ineffective teachers. This infamous report placed part of the blame on teacher preparation programs, leading to an era of unprecedented levels of politicization toward accountability measures in teacher preparation (Cochran-Smith et al., 2013). Throughout this time, national accrediting agencies played an important role in this increased politicization (Paige et al., 2002; 2003). However, the role of accreditation agencies changed in the early twenty-first century due to new federal programs that emphasized greater data-driven approaches to accountability. The Obama administration's Race to the Top (RttT) initiative and its successor in Every Student Succeeds Act (ESSA) (RttT, 2009; ESSA, 2015) greatly impacted the nation's two largest teacher preparation bodies—the Teacher Education Accreditation Council (TEAC) and the National Council for Accreditation of Teacher Education (NCATE). Ultimately, the ripple effect of these federal initiatives culminated in the 2013 merger of TEAC and NCATE. This merger resulted in one accreditation body to govern teacher preparation in the United States, the Council for the Accreditation of Educator Preparation (CAEP).

The hallmark of CAEP accreditation is its five standards that educator preparation programs must address with an array of different evidence sources. These five standards

address the themes of teacher candidate quality, partnerships with schools, content and pedagogical knowledge, impact of candidates on PK-12 learning, and the effectiveness of the organization's quality assurance system (CAEP, 2016). While CAEP encourages programs to utilize instruments created by their respective state educational agency to address these five standards, the burden falls on education institutions to create their own valid and reliable instruments if no such instrument exists in their state. Furthermore, the CAEP standards call for programs to address candidate dispositions on non-academic attributes, and to incorporate feedback from a multitude of stakeholders in institutional data-informed decision making (e.g. teacher candidates, student teaching supervisors, principals from partnership schools).

There are examples of initiatives by both state educational agencies and educator preparation programs to utilize surveys to understand perceptions of the effectiveness of programs. The Massachusetts Department of Elementary and Secondary Education (DESE), for example, is required by Massachusetts education regulations to utilize survey data to report on educator preparation quality in the Commonwealth (Mass. Gen. Laws c. 69, § 1B; c. 69, §§ 1J and 1K, as amended by St. 2010; c. 12, § 3; c. 71, §§ 38G, 38G ½; c. 76, §19). As a result, DESE created a collection of surveys that seeks input from four major stakeholders in educator preparation— teacher candidates, teacher completers (1 year after completion), classroom teachers who supervised teacher candidates and hiring principals of teacher candidates (DESE, 2017a). At the program level, initiatives like the Carnegie Corporation's Teachers for a New Era (TNE) in the mid-2000s spurred research conducted by educator preparation programs to reimagine how surveys are used for assessment (Ludlow, Pedulla, et al., 2008). One such example

was the research conducted by the TNE team at Boston College, where new surveys were developed to measure teacher candidate beliefs about teaching, their perceptions of program quality, and ultimately measures of change from entry to exit of the program (Enterline et al., 2008; Ludlow, Pedulla, et al., 2008; Reagan, 2011).

While these two examples only provide a limited perspective on the use of surveys in teacher education, it is clear the increased politicization of teacher education, coupled with new data-driven accreditation requirements mandated by CAEP, necessitates educator preparation programs to reimagine how they conduct internal assessment through surveys. It is with this perspective that I situate the development of this dissertation.

The surveys produced by both DESE and Boston College utilize traditional Likert-scale items (DESE, 2017a; Ludlow et al., 2008). However, recent studies in the field of instrument development have shown the effectiveness of scenario-based scale items to measure beliefs and attitudes (Chang, 2017; Ludlow et al., 2014; Ludlow, Anghel, et al., 2020; Reynolds, 2020). While the literature on scenario-based scales are limited, this method shows several advantages over traditional Likert-based scales (Ludlow et al., 2014; Ludlow, Baez-Cruz, et al., 2020). The scenario scale framework is based on an integration of the facet theory and sentence mapping procedures of Louis Guttman (Guttman, 1959; Guttman & Greenbaum, 1998) and the Rasch measurement principles as developed by Georg Rasch (Rasch, 1960/1980).

It is well documented that scales utilizing traditional Likert-based items seeking to measure psychological constructs are prone to skewed distributions and ceiling effects (Friborg et al., 2006). While the stakeholder surveys developed by DESE found the

instruments to be statistically valid and reliable (DESE, 2017a), results from their published 2015-2016 surveys demonstrated positively biased responses (DESE, 2016). Thus, given the successful work of Ludlow et al. (2014) and their utilization of scenario-based items, the decision was made to adopt that method and apply it to the field of teacher preparation.

As previously articulated, CAEP (2016) includes five standards to measure the effectiveness of an educator preparation program, including domains such as candidate quality and candidate's impact on PK-12 learning. However, it is well documented the most influential part of teacher preparation is the clinical component, often referred to as student teaching or the practicum experience (Cochran-Smith, 1991; Darling-Hammond, 2014; Evertson, 1990). Thus, given the importance of the practicum experience, the focus of this instrument will be to measure one aspect of that experience: supervision.

During the student teaching practicum experience, there exists a triad of individuals that are involved: a university supervisor, a K-12 classroom teacher (often referred to as the *cooperating teacher*), and the teacher candidate. This study was conducted in Massachusetts, where the terminology for the university supervisor and cooperating teacher is "program supervisor" and "supervising practitioner," respectively. However, this language is used solely in the actual survey instrument, and other discussion utilizes terminology most common in the literature. Lastly, while the cooperating teacher is a crucial, yet also traditionally under-appreciated figure in the practicum experience (Glickman & Bey, 1990), the focus of the supervision survey targets the university supervisor.

Purpose of the Study

The purpose of this study contributes to the field of teacher preparation research, particularly surrounding the clinical component as described above. The construction of this new instrument accomplishes multiple tasks. The instrument development process illuminates the most important qualities of supervision, more specifically of teacher candidates. Secondly, this study highlights and elevates the voice of teacher candidates, on how teacher candidates perceive their clinical experience and interaction with their supervisor.

Research Questions

The goals of this study led me to develop the following research questions:

1) *To what extent can a Rasch/Guttman Scenario (RGS) scale development approach be used successfully in the development of a university supervisor scale?'*
2) *To what extent does a RGS scale affect the quality of information gained as perceived by survey participants?*

Significance of the Study

As shown through the earlier discussion, there have been major impactful changes in the field of teacher preparation. One of the largest of these impactful changes revolves around teacher preparation accountability and its relationship with program accreditation. These specific data-driven reform efforts, including the consolidation of TEAC and NCATE, have resulted in a system of accountability that places high importance on the inclusion of a variety of stakeholders to provide multiple points of data. However, although organizations like CAEP stress the importance of "valid and reliable"

instruments for use as evidence to demonstrate program quality, there is a lack of such instruments available for teacher preparation programs to draw upon. Thus, the burden has been placed on these preparation programs to develop such instruments, all while taking time from other tasks, in addition to other reforms implemented during the last decade.

With these things in mind, there are several reasons why this study is significant. First, by developing a new instrument that is valid and reliable, preparation programs can meet accreditation standards by including this as an evidence source. However, beyond just meeting a requirement for accreditation purposes, this study contributes to the research on how teacher preparation programs can use stakeholder data as a means for program improvement. The instrument was constructed under the principles of Rasch-measurement, placing the construct of supervisor quality on a hierarchical continuum, with clear distinctions between high-performing supervisor traits and actions and low-performing supervisors, based on the literature.

Finally, this study builds upon the Rasch-Guttman-based scenario (RGS) scale development process. RGS scale development utilizes Rasch-measurement principles and facet theory to construct scenario-based scale items. While there has been growing research during the last few years (Antipkina & Ludlow, 2020; Chang, 2017; Chang et al., 2019; Ludlow et al., 2014; Ludlow et al., 2019; Ludlow, Anghel, et al., 2020; Ludlow, Baez-Cruz, et al., 2020; Reynolds, 2020), this study contributes new research to both the measurement development and teacher education fields.

CHAPTER TWO: LITERATURE REVIEW

In the following section, I review different parts of the literature that inform the rationale for conducting this dissertation. This literature review consists of three distinct portions: 1) a review of early beginnings of teacher education and supervision of teachers; 2) an examination of accreditation of teacher preparation programs and impact on the measurement of program quality; and 3) the development of the methodological approach that guides this dissertation, specifically that of Rasch measurement principles and Guttman's Facet Theory (i.e. the so-called Rasch/Guttman Scenario methodology [Ludlow, Baez-Cruz, et al., 2020]).

Review of American Teacher Education History

Although schools existed in all of the original 13 colonies, education and educator effectiveness reforms were most often spearheaded by New England. As early as 1647, precedent was set in Massachusetts for governmental supervision over schooling; it was in this year that a law was passed that required all towns to establish schools (Tanner & Tanner, 1987). The main purpose of this governmental supervision was to ensure that all children in Massachusetts schools received a basic education as mandated by the colonial government. It is important to note that this government supervision was not about individual teachers, rather it focused on the supervision of schools. Nonetheless, the tradition of governmental involvement with schools dates as far back as the history of the nation itself.

After the end of the American Revolution, when the United States shifted from a colonial form of government to a republic, other shifts in the supervision of schools took place. One important change was the emergence of the district system, as seen in

Massachusetts. In 1789, the Commonwealth of Massachusetts enacted a law that established district schools, in which a school committee was created for the purposes of hiring teachers and school visitations (Tanner & Tanner, 1987). Although these newly created district schools became increasingly popular (particularly in rural communities), many problems emerged. Tanner and Tanner note the most significant issue that arose during this era was the inability of many communities to adequately fund and perform supervision of teachers within their schools. This issue of inequity is not a surprise even in the modern age, rigorous supervision of teaching staff continues to be a burden in many communities. Wealthy communities possess the means to fund programs to support their citizens, while communities of lower socioeconomic status fundamentally lack the tax base to do so.

Under the leadership of State Representative James Carter, the Massachusetts House Committee on Education created the first state board of education in the United States (Tanner & Tanner, 1987). Despite his important place in this narrative, history more often remembers the individual tasked with the job as the first Secretary of the Massachusetts Board of Education—Horace Mann. Horace Mann served as Secretary of the Massachusetts Board of Education for twelve years, where he issued twelve annual reports detailing the problems of public education in the state, particularly the problem of districting. However, the shift to state control over public education was not Mann's only contribution; in fact, Mann's other work is what pertains most to this dissertation's discussion: teacher education.

Mann and other contemporaries became increasingly aware there was an important distinction between teaching school and keeping school (Tanner & Tanner,

1987). Around the time of the shift to state control over education, there was the creation of the first public school for the preparation of teachers. In 1839, the first normal school (as teacher preparation programs were originally called) was established in Lexington, Massachusetts. With Cyrus Peirce as the institution's first principal, the school still operates today in the form of Framingham State University (Framingham State University, 2021). The establishment of this *public* institution represented a major shift in the professionalism of teacher education.

The increasing professionalization of teachers and teacher education demanded the need for supervision, as was common practice in other industries. The jurisdiction of supervision for in-service teachers was initially performed by school superintendents, although this changed toward the end of the nineteenth century, as superintendents formed larger networks of schools (Glanz & Hazi, 2019). With expanding networks of schools, superintendents oversaw overwhelming numbers of teachers, resulting in a gradual shift to principals performing these acts (Tanner & Tanner, 1987). What was vital to these developments was the dialogue that took place between the newly emerging field of teacher education and normal schools, state departments of education, local school committees, districts, and supervising principals and superintendents. This early period in teacher education and supervision represented the beginning of the discussion of the *measurement* of teachers. How can you effectively and accurately *measure* what is a "good" teacher? As these discussions of educational supervision and measurement took place, the following several decades brought forth new scientific models of supervision outside the field of education.

Emergence of Clinical Supervision

In the modern field of teacher education, clinical supervision is commonly viewed as the primary form of supervision for teacher candidates. Prior to its emergence, Frederick Taylor (1911) pioneered the idea of scientific management, and how such management translated into greater workplace efficiency. It was Franklin Bobbitt (1913) who then brought the ideas of Taylor's scientific management and translated it to the educational context, in which supervision was situated in a highly bureaucratic structure and extremely control-oriented. Bobbitt's ideas were controversial, as he viewed education in schools as equivalent to production in factories (Glanz & Hazi, 2019). It was in response to Bobbitt's work that led to the movement of democratic supervision, notably the work of John Dewey (1929) and James Hocic (1920). Supervisors that engaged in democratic supervision employed scientific methods in their practice, but also cooperative problem-solving between teachers, instructors, and administrators to improve education (Pajak, 1993).

Modern clinical supervision takes inspiration from the Hunter Model, as well as the general professionalization of teacher education that had been taking place over the preceding decades (Hunter, 1973). Pavan (1985) notes the Hunter Model is viewed by many in the teacher education field as being synonymous with clinical supervision. Specifically, it is a system in which teachers are monitored and evaluated on classroom essential elements of instruction, and subsequently given critical feedback of these findings, reinforcing desired practices. It was during the late 1950s that one of the first concepts of clinical supervision emerged from Harvard's Master of Arts in Teaching program in conjunction with the Newton Public Schools (Tanner & Tanner, 1987). This

was a noted evolution in the history of pre-service teacher education as it captured a shift toward professionalization of the education of teachers.

There were other scholars during this time who sought to embody the type of clinical supervision seen in the medical field to preservice teacher education. James Conant (1963) called for clinical professors whose role would not be actively engaged in research or publishing. Rather, these clinical professors would hold joint appointments at the university and in school districts where they would be primarily involved with the supervision of teacher candidates at their field placements. Many medical institutions had these systems of supervision in place, and Conant argued that by implementing this in teacher education that the profession as a whole would rise in prestige.

Establishment of Accrediting Bodies in Teacher Education

In 1954, the National Council for Accreditation of Teacher Education (NCATE) was formed to become the sole agency responsible for teacher education in the United States (NCATE, 2008). Before its creation, the American Association of Colleges for Teacher Education (AACTE) was tasked with the accreditation of teacher education programs. Following this change, AACTE further shifted its focus to research in teacher education, to which it remains one of the premier organizations in that field today.

The formation of NCATE was a direct effort to hold teacher preparation programs accountable for their graduates. The push for NCATE's creation was the result of several different organizations, demonstrating a widespread commitment to teacher accountability. They included AACTE, the National Association of State Directors of Teacher Education and Certification (NASDTEC), the National Education Association (NEA), the Council of Chief State School Officers (CCSSO), and the National School

Boards Association (NSBA) (NCATE, 2008). While teacher preparation programs were not regulated by the federal government, there were clear directives for programs to comply with their standards. Like the CCSSO's modern push for the Common Core State Standards through federal programs like Race to the Top, programs were incentivized by external forces like state partnership programs to comply with NCATE accreditation (Henry et al., 2012). In some states, NCATE was viewed as a suitable alternative to a state-sponsored program review, and thus could function as evidence of sufficient program quality (NCATE, 2014).

As part of its standards for accreditation, NCATE asked teacher preparation programs to provide evidence of candidate assessment and performance in student teaching placements (NCATE, 2008). In the latest iteration of its standards before its 2013 merger with the Teacher Education Accreditation Council (TEAC) to form the Council for the Accreditation of Educator Preparation (CAEP), the standards explicitly called for teacher preparation programs to demonstrate evidence of "an assessment system that reflects the conceptual framework and professional and state standards" which serve as "evaluation measures to monitor candidate performance" (NCATE, 2008, p. 25). What emerges in these standards is a similar, but new word: evaluation.

Modern Accreditation and Impact on Teacher Education

The current landscape of teacher education is not only intensely politicized, but the systems of accountability and accreditation are as well. In the final iteration of its standards in 2008, NCATE continued to espouse the idea that teacher preparation programs should utilize their systems of supervision and evaluation of teacher candidates as forms of evidence to demonstrate program quality. The Teacher Education

Accreditation Council (TEAC)—the other major accrediting body for teacher preparation programs—placed greater importance on programs articulating program-specific claims of quality and providing evidence to substantiate those claims. In contrast, NCATE and its successor (with a near monopoly in the market)—Council for the Accreditation of Educator Preparation (CAEP)—are vastly more prescriptive of what is expected of programs to provide. At the heart of these new standards are assessments of teacher candidate performance in the field, utilizing "valid and reliable" instruments that demonstrate candidate quality (CAEP, 2016). It is then on teacher preparation programs to either develop or utilize state-sanctioned instruments that are valid and reliable.

Teacher Education Program Models – University-Based vs. Alternative

It is important to note that the term "teacher preparation program" itself encompasses a variety of different programmatic models. While university-based teacher preparation still represents the most common pathway for prospective teachers, the last 30 years has seen a resurgence of new "alternative" teacher preparation programs. However, the existence of alternative programs is not a new phenomenon, and in fact is as old as the field of teacher preparation itself. In addition to university-based programs, teachers were trained in a variety of contexts, including through school district programs, seminaries, community colleges, four-year schools, as well as specific programs to prepare teachers of color in segregated school districts (Fraser, 2007). Fraser notes "by 1914, virtually every city in the United States with a population of 300,000 or more and over 80% of those over 10,000 maintained their own teacher preparation program as part of the public school system" (p. 92). On the other hand, individuals from historically excluded gender, social class, and race/ethnic groups were prepared in alternative-like

programs. For example, between the first and second world wars of the twentieth century, children of Jewish immigrants in New York City were excluded from university preparation, and were trained at various programs in the city (Markowitz, 1993). Likewise, aspiring African-American teachers in the South were excluded from university preparation, and trained at Black educator preparation programs, some of which became four-year teacher colleges, like the institution in Montgomery, Alabama (Anderson, 1988). Finally, women were also barred from traditional university preparation, and trained under alternative measures. One such example were the women trained at the Keene State Normal School in New Hampshire, where they were taught in separate classrooms from their male counterparts, with gendered curricular tracks (Ogren, 2013). Thus, it is a complete misnomer that teacher preparation in the United States has solely revolved around the university pathway. In fact, in the history of American teacher education, there was only a brief period (1960-1990) where the university-based model held a true monopoly on teacher preparation (Zeichner, 2016).

In contrast to the alternative teacher preparation programs developed prior to 1960, alternative programs over the last 30 years have emerged under a different set of circumstances. The widespread critique of schools of education, most notably fueled by the report *A Nation at Risk* (1983), played a crucial role in the development of this new era of alternative programs (Zeichner, 2016). Whereas earlier alternative programs were often founded as reactions to discriminating toward certain populations (e.g. women, people of color, etc.), modern alternative programs were formed under different pretenses. These new alternative programs emerged as a direct response to a larger narrative that university-based programs lacked quality, which critics attributed to things

such as low standardized test scores and GPAs of teacher candidates, admissions policies, and program rigor (Cochran-Smith & Power, 2010). Therefore, critics argued that these low-quality programs were a "policy problem" that needed solving, by means of recruitment, certification, and preparation (Cochran-Smith, 2005). The most widely known alternative program in the United States is Teach for America (TFA). Founded in 1990, TFA's creation was an important milestone for alternative programs, with its specific goals of placing highly educated college graduates in high-poverty areas. Individuals selected for TFA participate in a six-week intensive training, and then are subsequently placed for a two-year commitment as a teacher in a high-need school (Cochran-Smith & Power, 2010). While TFA and other programs initially held strong partnerships with existing accredited colleges and university programs, changes led by the United States Department of Education allowed alternative programs like TFA greater flexibility to offer their programs independently from universities (Zeichner, 2016).

In addition to programs like TFA, growth of alternative programs coincided with the growth of national charter school networks (Zeichner, 2016). Both charter networks (e.g. Rocketship and Knowledge is Power (KIPP)), as well as individual charter schools (e.g. Match Charter School) founded their own independent teacher certification programs, specifically geared to train teachers for their own schools (Stitzlein & West, 2014). For example, the Match Charter School in Boston established the Sposoto Graduate School of Education, with an explicit goal to create "jaw-droppingly effective rookie teachers," while placing students in their "no excuses"-based school environment (Miller, 2017). Teacher candidates at Sposato enroll for a two-year commitment, in

which they spend the first year in a "residency" as a tutor or assistant teacher at a Boston area "no-excuse" school. If students successfully complete the first year of residency and are hired by a local school, they spend their second year as a full-time teacher of record, receiving semi-regular coaching, complete online courses, and receive evaluations from principals, external sources, and student achievement data (Miller, 2017). Ultimate completion of the two-year program results in a Masters of Effective Teaching (MET). These alternative programs have been dubbed as "new graduate schools of education" or NGSEs. They are unaffiliated with universities, provide programs lasting at least 9-12 months, and are sanctioned by their state to endorse teachers for licensure and award master-level degrees to program graduates (Cochran-Smith et al., 2020).

The Massachusetts Educator Preparation Context

As previously discussed, this study takes place in the Commonwealth of Massachusetts. It is important to note why that decision was made, and what implications that has for this study. First, the use of Massachusetts as the study locale is *not* intended to serve as a proxy for other states. This study concerns itself with teacher preparation in Massachusetts because of the impact that teacher education research and policy within the Commonwealth ultimately has on the field at a national level. For example, the passage of the Massachusetts Education Reform Law in 1993 represented a watershed moment in teacher education policy, notable for its inclusion of testing requirements. This is something that LeGeros (2013) notes as putting Massachusetts at the "cutting edge" of "licensure and standards-based reform" (p. 59). This law positioned Massachusetts ahead of federal policies, which shortly after would require teacher preparation programs to report the passage rates of their students on state assessments, as

mandated by the 1998 Reauthorization of the Higher Education Act. More recently, as documented by Power (2020), the piloting and ultimate rejection of the edTPA teacher candidate performance in the early 2010s by the Massachusetts Department of Elementary and Secondary Education (DESE) heralded the beginning of several states abandoning or scaling back their participation (Power, 2020). DESE opted to develop their own assessment tool for teacher candidates, the Candidate Assessment of Performance (CAP), implemented first in 2015. Most notably, this assessment tool aligns to the Commonwealth's existing educator evaluation system used for in-service teachers (DESE, 2019).

The impact of Massachusetts' teacher preparation programs are felt far beyond niche circles of teacher education policy, frequently extending into the national and global media ecosystems. The adoption and rejection of the edTPA by Massachusetts was covered at length in both local and national media, in which most media coverage was negatively slanted (Power, 2020). As Cochran-Smith and Dudley-Manning (2001) documented, the failure of 59% of the teacher education graduates in 1998 on a new state-mandated teacher test made national and international headlines, in what they dubbed "The Flunk Heard Round the World." Speaker of the Massachusetts House of Representatives Thomas Finneran made front page headlines, infamously remarking, "I'll tell you who won't be a teacher: The idiots who took that test and flunked so miserably—and, of course, the idiots who passed them" (Pressley, 1998a, p. 1). Chairman of the Massachusetts Board of Education John Silber offered an ultimatum to the Commonwealth's teacher preparation programs, stating, "I think it's time to put up or shut down" (Pressley, 1998b, p. 6). Other local and national editorials with similar

sentiments ultimately played an important role in what was ultimately included as part of the 1998 Reauthorization of the Higher Education Act (Fowler, 2001).

However, amidst these policy decisions both in Massachusetts and nationally, analyses of Massachusetts Tests for Educator Licensure (MTEL) data uncovered troubling results from a *psychometric* perspective, which demonstrated a "flawed test containing defective items" (Ludlow, 2001, p. 15). Revisions made to the MTEL tests in the early 2000s contributed to their continual use to this day, which recent research has shown that higher test MTEL scores are positive, statistically-significant predictors of in-service teacher evaluation ratings and student test scores (Cowan et al., 2020). As a result, the National Council on Teacher Quality (NCTQ) recently highlighted these research findings and recommended other states in the country follow Massachusetts' "careful approach" of verifying the validity and reliability of their testing instruments (NCTQ, 2021, p. 28). Given the influence Massachusetts' actions have toward shaping national issues of teacher preparation and policy, it was an appropriate site to situate this dissertation.

Finally, it is important to note that even within the Commonwealth, teacher preparation takes many forms. Of the 66 approved teacher preparation programs, 49 (74.24%) are categorized as "Higher Education Institutions," with the remaining programs labeled as "Educator Preparation Program Provider"—designations held by both TFA and Sposoto— (n=9, 13.64%), "Collaborative" (n=3, 4.55%), "Public School District" (n=2, 3.03%), "Charter District" (n=2, 3.03%), and "Private (Non-Public/Non-Special Ed) Schools" (n=1, 1.52%) (DESE, 2022). Furthermore, there is also great variability within the category of "Higher Education Institutions," as different

universities and colleges utilize varying programmatic structures and internal policies and procedures. Thus, despite that all programs fall under DESE's jurisdiction—and as a result—share many common requirements, each preparation program is unique in many aspects.

Different Roles in University-Based Teacher Education Programs

As articulated previously, traditional university-based teacher preparation programs must provide evidence of program quality and mechanisms for program improvement as part of state and national accreditation procedures. However, data used as evidence must be sourced from a variety of stakeholders involved in various roles of the preparation program. Below, I provide a list of eight of the most important figures in a teacher preparation program, a brief description of their role, and possible other names that role may be referred to as. The name of the role that is listed first will guide how it is referenced henceforth in this dissertation.

Teacher Candidate (TC): The individual seeking a degree/endorsement at the teacher preparation program; completes coursework and student teaching, most likely in a PK-12 setting; other names: *student teacher, pre-service teacher*

Cooperating Teacher: A PK-12 classroom teacher the teacher candidate is placed with during their student teaching practicum; is not employed by the teacher preparation program, although may receive a course stipend; works alongside the university supervisor in supervising/evaluating the TC; other names: *supervising practitioner, classroom teacher*

University Supervisor: An individual employed by the teacher preparation program in charge of supervising the teacher candidates' practicum experience;

may be full-time faculty with teaching responsibilities, but mostly are adjunct status focused solely on practicum supervision; other names: *program supervisor, clinical faculty, college supervisor*

Education Faculty: Individual employed by the teacher preparation program that teaches courses in content and/or educational pedagogy; may or may not be involved in the clinical experience; often are full-time faculty, sometimes with research and publication responsibilities in addition to advising and teaching; other names: *Professor of Education (of different ranks), university faculty, program faculty*

Practicum Director: Individual whose role at the teacher preparation program is to oversee the placement, supervision, and evaluation of TCs in their practicums; may or may not also be considered a Faculty member; may be tasked with issuing recommendations for a TCs application for a teaching license: other names: *student teaching director, licensure officer*

Program Alumna: Individuals that have completed a program at the teacher preparation program, usually within the last 5 years; may or may not be currently employed as an educator; other names: *in-service alumna*, *program graduate*

Hiring Principal: Individual who served as the principal at which a program alumna was hired for employment

PK-12 Partner: Individual employed by a public district or private school that acts as a liaison between their district/school and the teacher preparation program; may be the Principal, or another administrator; other names: *community liaison*

The above list is not exhaustive of all the roles that may exist within a teacher preparation program. However, for the purposes of this dissertation, these are the individuals that are most likely to be considered as a relevant "stakeholder" in program accreditation. While all of these roles may be valued as an important stakeholder, there are great discrepancies between roles in relation to their focus in existing research. Given that modern accreditation policies warrant programs to include data from various stakeholders, and that such data be developed from valid and reliable instruments, preparation programs are often put at a disadvantage when tasked to provide data and no existing instruments have been developed.

Much of the existing literature and developed instruments have focused mostly on the teacher candidate, primarily content/pedagogy coursework with education faculty, or their evaluation on a performance assessment during the practicum. Additionally, performance on state-mandated teaching tests is also often used as a measure of program quality. However, while there is a focus on the teacher candidate, their voice as a stakeholder is often muted. Furthermore, while course evaluations may provide an outlet for teacher candidates to give feedback to education faculty, there is a lack of instruments that have been developed to capture the quality of their supervision experience, particularly the university supervisor. Given the importance of the clinical experience, and the university supervisor's leading role, the lack of instruments and research is puzzling.

University Supervisor

As defined earlier, the university supervisor is an individual employed by the teacher preparation program, whose main role is to observe the student teacher,

collaborate with the cooperating teacher, and administer assessments from the preparation program. The role is one that has been a part of teacher preparation for decades, and while there is not an abundance of literature about this role, what has been long-made clear in research is the inherent complexity of this position. In one early example, Haines (1960) describes in extensive detail the role of the university supervisor:

> As liaison and public relations person, he helps to promote greater understanding of our participation in the preservice teacher education program. As a supervisory instructor, he assumes responsibility for encouraging the student teacher's continued professional growth and personal adjustment. As a co-worker in the public school he collaborates with the principal and cooperating teacher in improving the quality of preservice practical experience. The coordinated action influence and, in turn, are influenced by the participation of other key personnel who work closely with student teachers (p. 251)

While Haines (1960) certainly writes through a lens of admiration for the complexity of the role, the theme of the supervisor as an individual with many duties and responsibilities appears frequently. One of these recurring themes throughout the decades has been the supervisor's role of liaison between the preparation program and the K-12 school (Asplin & Marks, 2013; Briggs, 1963; Sharp, 1990). As the university supervisor is not with the teacher candidate every day, their influence on the candidate can often pale in comparison to the cooperating teacher, who works with the candidate daily (Marks, 2002). However, while the cooperating teacher may have a greater impact on the candidate's practice, the university supervisor often remains the individual with the most responsibility regarding *formal* observation, feedback, and evaluation (Morris, 1980).

It is crucially important to note that for the purposes of this dissertation, the term "university supervisor" refers to an individual that is *specifically* employed in that *role*, and not individuals that may engage in teacher candidate supervision as *part* of their role. There are many teacher preparation programs in which full-time faculty have

responsibilities both as instructors in content and pedagogy as well in the supervision of teacher candidates. However, the development of this instrument concerns itself with those individuals employed *solely* as university supervisors, a non-tenure track, part-time, adjunct role. These individuals are hired and evaluated under a completely different set of parameters than faculty who supervise, and this hiring and evaluation is conducted by entirely different sets of individuals (e.g. staff/administrators vs. faculty/department chairs/deans of faculty). Thus, this instrument, and therefore the study's participants, were recruited from teacher preparation programs that utilize adjunct university supervisors.

Significance of the University Supervisor

It is well documented throughout the literature that when in-service teachers are asked which part of their teacher preparation program was most influential, the field experience consistently earns the top factor (Clifford & Guthrie, 1990; Guyton & McIntyre, 1990; Wilson et al., 2002). However, despite a wide consensus of the importance of student teaching, and with a robust field of research on it, there is little research on university supervisors (Steadman & Brown, 2011). Enz et al. (1996) described the state of research on university supervisors as "relatively sparse." (p. 132), and the field of research has not grown much in the years since. This presents an odd paradox, as the university supervisor occupies a significant role in what is considered a vital part of teacher education, yet as a group, they are understudied. Darling-Hammond (2014) posits that the quality and intensity of supervision, and the evaluation tools used during student teaching, are important elements of teacher learning. Given the university

supervisor is the individual most responsible for implementing these elements, it is important to understand their practices.

As Steadman and Brown (2011) found in their qualitative case study of university supervisors, there are great inconsistencies in how supervision is enacted, even within the same education department. Inconsistency and lack of clarity regarding role expectations is something common among university supervisors (Ganser, 1996) and do a great disservice to the education of the student teacher. Regarding the triad of student teacher-university supervisor-cooperating teacher, Johnson and Napper-Owen (2011) highlight how crucial cohesion is, writing that "the formation of a collaborative group is essential for a successful and positive experience" (p. 45). When a positive relationship is formed between university supervisor and student teacher, it increases the likelihood of student teacher enacting on their supervisor's feedback and implementing changes to their teaching practice (Asplin & Marks, 2013).

The university supervisor position is unique in the larger context of teacher education. While they sometimes may serve as a faculty member in the school of education, by and large they are composed of three types of individuals, what Cochran-Smith (2003) calls "*on the side,*" "*in the middle*," and "*at the end*" (p. 14). The first group "on the side" refers to graduate students studying education, who take this position as a side-job or graduate assistantship. The second group "in the middle" refers to teachers that have temporarily stepped away from full-time K-12 teaching (e.g. maternity leave), and take this position with an intention to return to K-12 someday. The final group "at the end" is perhaps the largest group, which is comprised of retired teachers that have stepped away from K-12 teaching, and take this position as a retirement job.

With a diverse set of individuals who serve as university supervisors, their role is often misunderstood. While employed by a university or preparation program, university supervisors are not viewed as peers to tenure-track education faculty (Anderson et al., 1992). Steadman and Brown (2011) hypothesize one reason for the lack of research on university supervisors may partially stem from the tension between faculty and university supervisor. This tension is attributed to a rift over which is more important, the theory of teacher education (taught by faculty) or the application of theory (supervision and field experience). Additionally, while the university supervisor's work environment is the K-12 classroom, they are consistently operating as an "outsider." They are university-employed but not viewed as university educators. They work closely with K-12 classroom teachers, but they operate as a guest in that teacher's classroom. As a result, the university supervisor's purpose has been viewed as primarily to serve as a bridge between the preparation program and K-12 classroom (Koerner & Rust, 2002).

There is not unanimous consensus regarding the importance of the university supervisor. Given the lack of clarity over roles and expectations, some researchers view the university supervisor as superfluous in the overall context of teacher education (Metcalf, 1991; Zeichner & Gore, 1990). However, the overwhelming interpretation of this lack of role clarity is not the irrelevancy of the university supervisor, but of the tremendous potential the role can serve when enacted well (Slick, 1997; Steadman & Brown, 2011; Potter et al., 2016). Given the universally accepted importance of the clinical experience, in tandem with a recognized untapped potential of the university supervisor role, it is important for the field to better understand this role, and what attributes makes a quality supervisor. In order to understand this construct of university

supervisor quality, one must have evidence and data to support any claims. Given a lack of existing tools, one approach is to use the principles of scale development, so that we can *measure* our construct through scale items. It is from this that this dissertation's purpose arises, to develop a scale measuring the construct of the quality of the university supervisor in the student teaching experience.

Scenario-Based Survey Items

As described in several sections throughout this dissertation, surveys play an important role in the evaluation of teacher education programs, most significantly in accreditation. However, not only is there a lack of instrumentation targeting the quality of university supervisors (Cuencaa et al., 2011), even those existing Likert-scale instruments in the teacher preparation field merit serious revisiting. In recent years, as a response to biased traditional Likert-scale survey items, a new approach utilizes both the principles of Rasch measurement principles and Louis Guttman's Facet Theory. This approach was developed by Ludlow and applied by former and current students.

The traditional scale development process, when concerned with the measurement of emotional/psychological traits and/or beliefs, is rather formulaic and consistent across applications (Krosnick & Presser, 2010). Constructed scale items tend to be short in length, which has been common practice to help the individual with easier scale readability. When individuals are better equipped to answer scale items without added irrelevant difficulty of reading items, this helps reduce the measurement error of the scale. Lower overall measurement error results in a more accurate measure of whatever construct of interest is being measured. Additionally, the shorter length of items allows

for more scale items to be included in any given scale. This supports more statistical exploration of correlation patterns between items.

However, while there are clear benefits to these traditional short Likert-response items, there are issues that arise during the interpretation stage of scale development. One of these problems manifests itself through one of the core tenets of research and measurement—validity. As described by Messick (1989), validity is defined as an overall evaluative judgment, as it relates to the "adequacy" and "appropriateness" of the "interpretations" and "actions" as based on a test or set of scale items. When tests or scales have strong validity, then greater faith can be placed in subsequent interpretations or actions based on the psychometric analyses conducted on data. Yet, many traditionally constructed scales, mostly utilizing Likert-items, often fail to capture deeper complexities of constructs (Ludlow, Baez-Cruz, et al., 2020). This failure of items to authentically measure the true scope of a construct is a direct threat to validity.

Issues with traditional scale development also extend beyond the length and number of items. Most often, the Likert response categories exist on a continuum of "Strongly Disagree" to "Strongly Agree," or "Never" to "Always." However, scale items are sometimes inappropriately paired with incompatible category choices, particularly when part of longitudinal studies. For example in 2017, DESE changed its response categories on their stakeholder surveys from the options "Strongly Agree, Agree, Neither Agree Nor Disagree, Disagree, Strongly Disagree" to "Agree, Somewhat Agree, Neither Agree Nor Disagree, Somewhat Disagree, Disagree." (DESE, 2017a). These surveys play an integral role in the evaluation of teacher preparation programs in the Commonwealth, which utilize survey results across several years during a program's review cycle.

Programs reviewed during this cycle were scored partially on a previously established criterion, which was the percentage of stakeholder responses that answered "Agree" or higher. However, in surveys previous to 2016, Agree was the second-highest option, whereas in 2017 and beyond, it was the highest response option. The rationale behind this change was that DESE argued the response option "Somewhat Agree" implied doubt, and thus its scores could not be aggregated with those that answered "Agree," even though this "doubt" was on the positive side of the scale spectrum. Interesting enough, when the Educator Preparation team publishes their annual review, which includes feedback from preparation providers in the Commonwealth about the agency, their reports aggregate both Agree and Somewhat Agree together (DESE, 2017c).

The spacing of response categories (i.e. the "jump" from Strongly Agree to Agree) can also manifest itself in problematic ways in various scales. For example, the Higher Education Research Institute (HERI) freshman year student survey is one of the most widely used scales in higher education, administrated annually to tens of thousands of students across the United States. While most scales contain five response categories, utilizing something similar to "Strongly Agree, Agree, Neither Agree Nor Disagree, Disagree, Strongly Disagree," this survey uses only three options, "Frequently, Occasionally, Not At All." (HERI, 2020). Not only does the fewer number of categories lend itself to more potential psychometric instabilities, but the continuum of response categories is uneven. The bottom response category "Not at all" is an extreme response option, whereas its counterpart of the high end of the spectrum "Frequently" is not extreme (unlike *per se*, "All the time"). The lack of response options as well as their

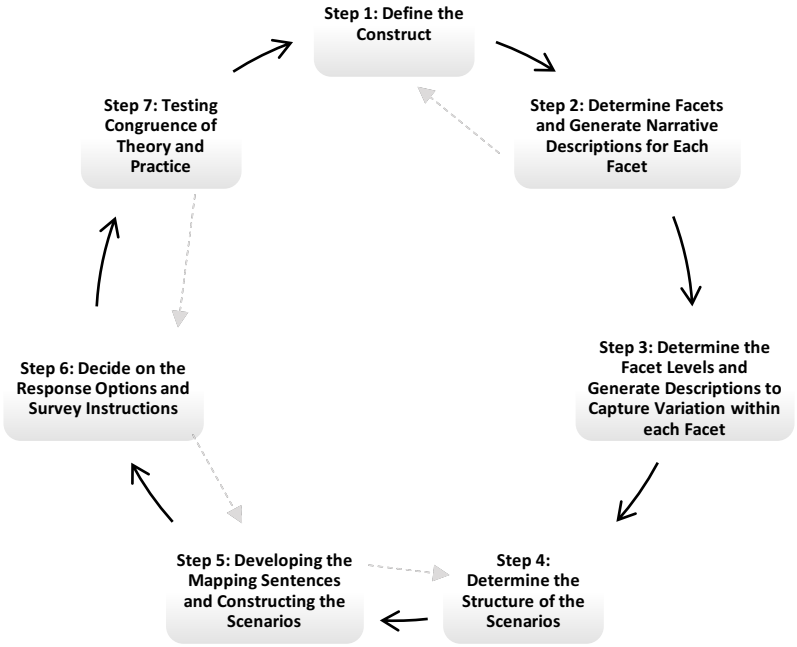
severity can lead to greater measurement error and misfit scale items, that is, when observed responses greatly deviate from their expected responses.

There is also much discussed in the literature about issues related to the middle response option, often labeled “Neither agree nor disagree,” “Unsure,” or some other phrasing. This middle response category is one that is often misunderstood or inappropriately used throughout many scales (Truebner, 2019). At times, while intended to be the midpoint response category, this option (particularly with attitudinal scales), functions as a response option for individuals that are confused and/or ambivalent to the item (Truebner, 2019). This inappropriate functioning of the midpoint category contributes to increased measurement error. However, other studies have shown that the omission of a midpoint response option is also problematic, and can contribute toward increased measurement error (Bishop, 1987; Kalton et al., 1980). Without this option, individuals may feel compelled to provide answers that are more extreme than their true beliefs. Thus, it is important to include a midpoint response option, but it must function properly and be clear to the individual what its value means.

The Rasch-Guttman-based scenario (RGS) scales developed over the last decade have demonstrated promising results in combatting the issues found in traditional scale development. Ludlow, Baez-Cruz, et al. (2020) offer a framework for constructing RGS scales, a process that is comprised of seven stages. In the following discussion, I present my processes that utilized this framework put forth by Ludlow, Baez-Cruz, et al. (2020), and displayed in Figure 2.1.

Figure 2.1

Diagram of the RGS scenario development process (Ludlow, Baez-Cruz, et al., 2020, p. 365)



Seven Steps of Developing RGS Scales

Step 1: Define the Construct

Displayed in Figure 1, the first step in constructing RGS scales is to "define the construct" (Ludlow, Baez-Cruz, et al., 2020). As articulated earlier, the goal of this dissertation is to develop a scale measuring the quality of the university supervisor in the student teaching triad. The construct of interest for this scale is "the quality of student teacher supervision." While perhaps a seemingly simple step in this process, it is perhaps the most crucial step to clearly define, as guided the remaining steps of this dissertation.

This construct has three main components that I break down in the following section: 1) what attribute is being measured (i.e. "quality"); 2) whose attribute is being

measured (i.e. the "supervisor"; and 3) what context is this scale occurring under (i.e. the "student teaching" experience). The first of these three is "quality," which is an important distinction from "quantity." If this scale were to measure the "quantity" of supervision that a teacher candidate received during their student teaching experience, then the development of a RGS scale would be inherently unnecessary. In lieu of an RGS scale, one could merely develop a checklist of supervisor tasks (e.g. # of lesson plans reviewed by supervisor, # of site visits, etc.), and at the end, tabulate the sum of all the tasks. However, in order to measure *quality*, a more subjective attribute, it was vital to employ more in-depth research and hypotheses as to what *quality* truly means. The second component is whom the attribute of interest (i.e. "quality") is directed at; in this case, it is the supervisor. Although there is a great abundance of literature throughout the field of teacher education about measuring the quality of *student teachers*, this instrument focused not on *their* performance quality, but of their *supervisor*. This is an important distinction, as the teacher candidates were the survey respondents, specifically regarding their perceptions of their university supervisor's quality, and not their own. Finally, the third component situates what field the scale takes place within, which is the student teaching practicum. There are existing scales concerning supervisor quality in pre-service trainee fields, such as nursing (Saarikoski, 2014) and counseling psychology (Cook et al., 2018; Winstanley & White, 2014); however, while there may be similarities across fields, it is important for the scale's structure and development to be explicitly defined as occurring in the student teaching practicum.

Step 2: Determine Facets and Generate Narrative Descriptions for Each Facet

With a clearly defined construct, the second step of RGS scale development is to break down that overlying construct and deconstruct it into *facets*. As Ludlow, Baez-Cruz, et al. (2020) note, the identification of "main characteristics, principles, or characteristics," is that in sum equal *facets* (Ludlow, Baez-Cruz, et al., 2020, p. 366). While facets of a construct will inherently by their nature be related to one another, it is important each facet be clearly distinguishable from one another. Too much overlap between facets creates issues down the line during scenario writing, and in the analyses and interpretation of results.

While RGS scale development is guided by robust quantitative processes, such as the adherence to the principles of Rasch measurement, it is important to note ambiguity exists regarding the *number* of facets in various RGS scales. Simply put, there is no one set number of facets per RGS scale, nor a formula designed to produce the "correct" number of facets. As Ludlow, Baez-Cruz, et al. (2020) reflect on the numerous successful RGS scales they have created, their principle is that facet identification ends when the scale developer(s) concludes that the construct is comprehensively represented. The scale developer(s) must ultimately use their expertise and thorough examination of the literature to justify which and how many facets to use to define the construct.

Step 3: Determine the Facet Levels and Generate Descriptions to Capture Variation within each Facet

The third step is to determine the levels of variation that occur within a given facet, and to attach labels to these levels. As Ludlow, Baez-Cruz, et al. (2020) describe, all their RGS scales focused on low, moderate, and high levels of each facet, or *struts*. A

set number of three struts is ideal for multiple reasons, both pragmatic and theoretical. Pragmatically, it forces the scale developer to theorize levels of graduation at the onset; if a facet fails to clearly show three levels of graduation, then latter steps of scenario building will almost certainly fail given the need will arise to differentiate the facet at an even finer level. Furthermore, this process helps align with theory, and helps improves the proof of concept of a given construct, and minimizes inconsistencies of facet descriptions (Ludlow et al., 2014).

Following the creation of these levels, they are then each assigned a corresponding numerical value. The numerical values for each strut are then summed alongside other facet scores to provide a *structuple* level for a scenario (Ludlow, Baez-Cruz, et al., 2020). The values assigned to struts are consistent across most RGS scales, with a value of three to facets written at a high level, a value of two to the medium level, and a value of one at the low level (Chang, 2017; Ludlow et al., 2014; Reynolds, 2020). The total numerical score will vary based on the number of facets, something that varies across RGS scales. For example, where the scales of Ludlow et al. (2014) and Reynolds (2020) were comprised of four facets, the scale of Chang (2017) included six facets. For a scale of four facets, a scenario scored at the high level of all four facets would have a total score of 12, whereas a scenario scored at the low level on all four facets would have a total score of 4. The process would look nearly identical for a scale such as Chang's (2017), albeit the highest score would be 18 (i.e. 3x6) and the lowest score of 6 (i.e. 1x 6).

This numerical coding is crucial to the RGS scale development process, as it lays the foundation for the creation of scenarios. The number of scenarios ultimately created

will equal the number of struts multiplied by the number of facets, minus the number of overall struts. In the case of a RGS scale with three struts and four facets, this would lead to a scale with nine scenarios. As Ludlow, Baez-Cruz, et al. (2020) note, the RGS approach is different than Roskam and Broers (1996) and Randall and Engelhard (2010), who create all possible combinations of facets and struts. If the RGS method were to employ that approach, then a scale with three struts and four facets would consist of 81 scenarios (3^3) and a three strut/six facet scale would consist of 729 (3^6) scenarios. Not only would this be wildly impractical in application, it would almost certainly include scenario combinations inconsistent with existing theory of how a construct and its facet would manifest.

Step 4: Determine the Structure of the Scenarios

As described in the previous section, it would be both impractical and illogical to construct every potential facet and strut, and thus step four of the RGS scale development process is to select the combinations that will take form in the scale. One approach in this fourth step is to initially construct scenarios at the extreme levels, that is, at both the high and low ends of each facet and strut. For a four-facet, three strut scale, this would result in one scenario written with four facets at the highest level, and a second scenario with four facets written at a low level. This approach is useful as it clearly sets the bounds of the scale, and subsequently sets the stage for the creation of mid-level items. While some scales may have a rather formulaic structure in their overall composition, like an equal number of high/medium/low scenarios (e.g. the PEPS scale, Ludlow et al., 2014), it is crucially important the scale developer can establish a "proof of concept," meaning that the scale is grounded in theory (Ludlow, Baez-Cruz, et al., 2020, p. 369).

Ludlow, Baez-Cruz, et al. (2020) suggest that during this fourth step, the scale developer consider four different points regarding the construction of various facet/strut combinations:

1. Does the literature/existing empirical evidence support the facet combinations?
2. Do the facet combinations fully cover the construct domain?
3. Are the selected facet combinations logical and reasonable?
4. Do the facet combinations result in a total number of scenarios that is feasible? (p. 369)

These four points provide the scale developer with a solid foundation for creating scenarios, forcing the answering of crucial questions. Per the first point, any scenario constructed must be based on evidence backed by the literature, or other forms of evidence (e.g. focus-group data, personal experience, etc.). If a scenario's composition cannot be supported with forms of evidence, this is a direct threat to validity. The second point is also important, as the purpose of any scale development is to fully capture a construct through the use of questions and/or items (Hambleton & Cook, 1977). If the scale developer cannot argue their scenario structure encompasses the entirety of the construct, then they must continue to revise it until it does. While the third point may seem like an unnecessary inclusion, given its inherent obviousness (i.e. does your scale make sense?), it is important to consider. Aside from the extreme scenarios, the combinations of facets and struts between those extremes ultimately becomes the judgment of the scale developer. While one *could* write a scenario with three facets at the high strut level, and one facet at the low strut level, it most likely would not make *logical*

sense. Lastly, as articulated earlier, there could be a wide range of potential scenarios if all combinations were written. From a practical standpoint, it is important to consider the burden on the scale-taker, and to construct a scale of reasonable length.

Step 5: Developing the Mapping Sentences and Constructing the Scenarios

With an established set of struts per facet, the fifth step of the RGS process is to construct mapping sentences. Ludlow, Baez-Cruz, et al. (2020) writes that this process is both formal (i.e. elements from the struts are utilized) and informal (i.e. phrases are interspersed linking struts to one another). This process draws upon Borg and Shye (1995), who provide a framework for mapping sentences. In this format, facets and their strut levels are presented in parentheses, and the names of the facets are then placed onto of the struts within the mapping sentence. Subsequently, the informal linking phrases are then italicized, to indicate its purpose as a transitionary phrase (Ludlow, Baez-Cruz, et al., 2020). It is important to note that most RGS scales utilize "unobtrusive facetization," meaning that facet level descriptions are *not* repeated, nor are they explicitly stated as a "high" or "low" strut within a scenario (Borg & Shye, 1995, p. 34). As a result, this necessitates the need for constructing scenarios with a wide range of synonyms and similar (but unique) phrasing.

In practice, the sentence mapping directly sets up the structure for scenario creation. As demonstrated in *Figure 2*, Reynolds' (2020) physical accessibility scale employed the following sentence mapping, which she used to generate her scenarios.

Figure 2.2

Reynolds' (2020) Physical Accessibility Scale Sentence Map

Mapping Sentence 1. Physical Accessibility

Professor X has

Facet M: Arranged Meetings
(high [sufficient & flexible])
(medium)
(low [insufficient & inflexible])

times for arranged meetings with students. S/he is

Facet C: Chance Encounters
(high [a constant presence])
(medium)
(low [never seen])

around campus.

Professor X's responses to email are

Facet E: Email
(high [timely & consistent])
(medium)
(low [untimely & inconsistent])

As Figure 2.1 demonstrates, the italicized words represent the informal transitionary phrases, whereas the facets are separately defined, and their multiple struts of high/medium/low are presented as well. It becomes clear how one would develop a scenario given this sentence mapping structure, starting from the upper left corner, and working across and downward. While some RGS scales like Reynolds (2020) vary the order in which facets appear across their scenarios, other RGS scales maintain a consistent structure throughout the entirety of the scale.

Step 6: Decide on the Response Options and Survey Instructions

One of the greatest challenges in scale development is constructing items that are accessible and relatable to the individual answering the items. Too often scale items are constructed in such a manner that is either confusing and/or feels too removed from the reality of the scale-taker. All four RGS scales described in the article by Ludlow, Baez-

Cruz, et al. (2020) employ a comparative scenario response format. Whereas in a traditional scale respondents respond to statements on an agree-to-disagree spectrum, this approach asks participants to compare their lived experiences to the experience described in the scenario.

In practice, this comparative approach provides for a slightly more complicated user experience; however, it also has been shown to aid in reducing social desirability bias (Ludlow et al., 2014). As an example, the instructions on Reynolds' (2020) physical accessibility scale asks respondents to compare their faculty member's performance to the one described in the scenario. On this scale, the comparative response options from high to low were as follows:

They are much more accessible than Professor X

They are a little more accessible than Professor X

They and Professor C are about the same

Professor C is a little more accessible than them.

Professor C is much more accessible than them.

While some RGS scales, like the Productive Engagement Portfolio (PEPS) scale (Ludlow et al., 2014), may use a mix of both positive (e.g. *more* engaged) and negative (e.g. *less* engaged) terminology, other RGS scales remove negative terminology (like the physical accessibility scale above) and reverse the ordering of the comparisons. In practice, this has been shown to help reduce social desirability bias, as respondents may be reluctant to admit to being "less than." (Antipkina & Ludlow, 2020).

Step 7: Testing Congruence of Theory and Practice

The seventh and final step of RGS scale development is to take the set of generated scenarios and administer them. Like all scale development processes, the administration step will likely take multiple different forms. Scenarios may be presented informally with colleagues or with focus groups, and workshopped to revise words and phrases that may be confusing or need editing. Informal pre-pilot studies with a smaller group of respondents may take place, before a more formal larger scale pilot study. The process of RGS scale development process is iterative, as revisions will occur as results are analyzed, and subsequent changes are made to address any issues that arose during psychometric analysis. There is no set number of pilot administrations that occur before a RGS scale is finalized; the process is only complete when the psychometric properties of the data convey a stable result, where scenario items are evenly distributed and spaced in their difficulties, and their overall ordering fits the original hypothesized structure.

The manner in which RGS scales are analyzed is contained in its name, which is under Rasch measurement principles. In the following section, I provide an overview of Rasch measurement.

Rasch Measurement Theory

The core measurement principles that are embodied by these scenario-based survey items are Rasch measurement principles. For the purposes of this dissertation, I use the Rasch Rating Scale model, as described by Wright and Masters (1982), and utilized in other scenario-based survey scale items (Antipkina & Ludlow, 2020; Chang, 2017; Ludlow et al., 2014; Reynolds, 2020). Thus, it is important to first understand what the principles of Rasch measurement are, and how they will inform this dissertation.

Rasch Measurement Principles

There are several key principles that embody Rasch measurement. The work of Georg Rasch (1960/1980) resulted in a multitude of research in the field, spanning several decades, published in thousands of journal articles, and taught across hundreds of universities across the globe. However, despite its prevalence in the literature, Ludlow et al. (2014) that provides one of (if not the first) published documentations of what *exactly* are the defining principles of Rasch measurement.

The first of these principles as defined by Ludlow et al. (2014) is that the variable or construct being measured is unidimensional. This means that whatever is being measured, the measurement tool captures a singular dimension of that. The second principle is that the items should exist on a spectrum of progressive difficulty. This means there should be relatively "easy" items, "moderately difficult" items, and "difficult" items. This "difficulty" term can be applied toward all types of scales, from academic tests to physical health examinations to mental health surveys. For the latter two, difficulty is viewed as more difficult to perform a task (i.e. physical health exam) or more difficult to endorse agreement with a statement (i.e. mental health survey).

The third principle is that the difficulty progression of the items on the scale should be spread across it in an even manner (Ludlow et al., 2014). One metaphor that captures this idea is to imagine the scale as a ladder, with each item represented by an evenly spaced rung, with the spaces between rungs remaining equal throughout (i.e. spaces between bottom rungs are not smaller than spaces between rungs at the top). Building upon this even spread, the fourth principle is that each individual item has an

equal discrimination, meaning that each item equally correlates with the scale's total score.

The fifth principle articulated by Ludlow et al. (2014) is the independence of items from one another. This entails that answers to one item are not contingent on the answers of other items. Finally, the sixth principle is that items should be analyzed and "weeding" or discarding of poorly fitting items should occur (Rasch, 1960/1980, p. 125). This is done so that all items fit together well, both from a perspective of fitting the data well, and any hypothesized theory of how the data were to emerge.

Rasch Measurement Models

There are several types of Rasch models other than the Rating Scale model, which include dichotomous and partial credit models, both of which are one-parameter logistic model (1PL) Item-Response Theory (IRT) models. Beyond the 1-PL there are both the two-parameter logistic model (2PL) and 3-parameter logistic model (3PL), each of which have important similarities and differences; however, for the purposes of this dissertation, discussion will focus on the 1PL model. The formula for the Rasch rating scale model is represented by the following formula below:

$$\pi_{nix} = \frac{e^{\sum_{j=0}^{x_{ni}}[\beta_n-(\delta_i+\tau_j)]}}{\sum_{k=0}^{m} e^{\sum_{j=0}^{k}[\beta_n-(\delta_i+\tau_j)]}}$$

On the left side of the equation, π_{nix} is the probability of person *n* responding to item *i* in category x. In this case, category *x* is represented by one of the response options provided to the student. The difficulty or location of the item *i* on the variable is represented by δ_i, and τ_j considers the difficulty of moving from one response category

to the next higher one. Lastly, x equals 0, 1, …., m where m is the maximum category score possible.

There are several reasons for choosing the Rating Scale Model over the partial credit model. The partial credit model will always fit the data better because it has more degrees of freedom, and thus will have a smaller log-likelihood chi-square value; however, for this scale, I use a Likert scale of Strongly Agree to Strongly Disagree. Wright and Masters (1982) found that the rating scale model performs particularly well when the response options are ordered, such as Strongly Agree to Strongly Disagree. The response categories remain fixed, as Wright and Masters (1982) state that "the relative difficulties of the steps in each item should not vary from item to item" (p. 48).

The 1-PL Rasch model only accounts for item difficulty in its equation, the two-parameter model includes item difficulty and discrimination, and finally the three-parameter model includes item difficulty, discrimination and a sort of pseudo-guessing parameter (Ludlow, Enterline, & Cochran-Smith, 2008). The Rasch model is the simplest in its statistical form of these models in that it does not include item discrimination and a guessing parameter, but the purpose of the Rasch model is to locate an individual's ability directly on a unidimensional scale, assuming the only thing influencing the result is their ability. 3-PL IRT models serve a purpose in large scale assessments like TIMMS and PIRLS, but this instrument is most appropriately developed for a 1-PL Rasch model because it measures beliefs, and guessing is not a part of this measurement. Furthermore, as discussed prior, one of the primary assumptions of the Rasch model is that all the items on the scale discriminate equally (Hambleton et al., 1991).

CHAPTER THREE: METHODOLOGY

In this chapter, I discuss the methodology enacted to answer the research questions of this dissertation. As part of this discussion, I describe the overall research design, which is comprised of four stages. After this, I describe in detail how the RGS scale development across unfolded, across its seven steps. Finally, I provide the rationale for the statistical model and procedures I chose for analyses.

Research Design

This dissertation's goal is to develop an instrument that measures the quality of the university supervisor during student teaching. This process mirrors that of how conventional instrument development occurs. The process that I employ can be broken down into four unique stages: 1) define the construct of interest 2) develop items, with informal workshopping with peers and content experts, and informal testing 3) implement a formal pilot study, and lastly 4) execute a final administration.

Participants

It is important to note the context within which this study takes place, which is in traditional, university-based teacher preparation program, whose programmatic structures utilize part-time, adjunct non-tenure track university supervisors. As described earlier, modern alternative programs are structured in vastly different manners than their university-based counterparts. Additionally, unlike university-based preparation programs, alternative programs generally do not pursue national accreditation, of which the development of this scale is purposefully designed for. Moreover, this study was designed for university-based programs that utilize adjunct, non-tenure track university supervisors, as compared to full-time faculty who teach, publish, and supervise. Non-

faculty university supervisors are hired, evaluated, and supervised in a completely different manner than full time faculty who supervise teacher candidates. Adjunct university supervisors are generally hired and supervised by a Practicum Director, or comparable staff position. In contrast, full time faculty are hired, retained, promoted, and evaluated under entirely different conditions, and report to senior-level administrators (e.g. department head, Dean of Faculty.) The programmatic structure of adjunct university supervisors exists across the spectrum of university-based programs, from large and small institutions, public and private, Research I (R1) and non-R1, or undergraduate/graduate only. The important defining characteristic is the adjunct status of the university supervisor, as the instrument is designed for the Practicum Director.

The participants of this study are individuals who completed a student teaching full practicum experience in the previous three years. Participants were recruited based on their completion of a full practicum from a university-based program that utilizes adjunct university supervisors. These participants were recruited at both the undergraduate and graduate levels. Individuals were recruited across all disciplines, including elementary education, secondary education, moderate and severe special needs, and reading.

Sampling

It is without question the importance of obtaining a sufficient sample size during the instrument development process. There are however, a variety of opinions and positions on what exactly constitutes a sufficient sample size. One commonly cited figure is Crocker and Algina's (1986) recommendation of a minimum of 200 participants for item development. Other recommendations stray from a minimum number of

participants, instead focusing on how many items are on the scale. For example, Nunnally (1967) posits that sample size should roughly equal somewhere between five and ten times the number of items. On the other end of the spectrum, it is also important to not *over*-sample. As Hair et al. (2010) describe, researchers should be careful with sample sizes over 400 participants, as it can possibly result in an over-powered study. As statistical power is a function influenced by sample size, too large of a sample may lead to statistically significant results, even when differences are functionally meaningless. Thus, elected to obtain a sample based on the number of items. I utilize nine scenario items, with five response categories. This resulted in a sample size of 76 and 61 participants, for the pilot and full administrations, respectively.

However, it is also important to reiterate that this scale is constructed under Rasch measurement principles, which offers advantages over classical test theory in regard to sample size. Wright (1977) stipulates that a sample size of just 100 is appropriate for Rasch analyses, half of what Crocker and Algina (1986) recommend. Nevertheless, successful instruments have been developed with samples under 100, including instruments developed under the RGS framework. Chang (2017) developed a psychometrically stable RGS instrument with 57 participants. Ultimately, 61 participants were procured for this survey instrument.

Rasch/Guttman Scenario Scale Development

As described earlier, and outlined by Ludlow, Baez-Cruz, et al. (2020), there are seven steps in RGS scale development. In the following section, I describe my procedures for these seven steps.

Step 1: Define the Construct

As articulated throughout, the construct of interest for this dissertation is the quality of the university supervisor. The goal of this dissertation is to create scenarios along a continuum, which captures this construct at the high end of quality (i.e. the optimal supervisor), middle (i.e. an average supervisor), and the low end (i.e. the worst supervisor).

Step 2: Determine Facets and Generate Narrative Descriptions for Each Facet

After thorough examination of the literature to properly define my construct of university supervisor quality, the next step was to determine what and how many facets existed within this construct. As previously described, there is no pre-established number of facets, and thus it becomes the expert judgment of the scale developer as to the appropriate number, and whether the construct is fully defined by such facets. In consultation with the literature, workshopping with peers and teacher educators, and through my own lived experiences working with student teachers and university supervisors for over five years, I determined four unique facets comprise the construct of university supervisor quality. I briefly describe each facet in the section below:

Resourcefulness: This first facet emerged clearly throughout the literature on university supervisors. The Association for Student Teaching (1964) describes this element of the university supervisor role as being a "dispenser of information" (p. 13), in that they are expected by both the student teacher, cooperating teacher, and other K-12 staff (e.g. school principal) to be a "walking encyclopedia" of information about the student teaching experience, and policies of the university (Association for Student Teaching, 1964, p. 13). Furthermore, the university supervisor acts as liaison between the

university and school, and in articulating the program's vision and goals to the student teacher (Byrd & Fogelman, 2012; Cuenca, 2010). Zimpher et al. (1980) describes the university supervisor as the "watchdog" of the triad to ensure the student teacher completes all the necessary requirements of the practicum (p. 14). Inconsistencies in the level of knowledge base of supervisors is directly tied with reported student perceptions of satisfaction with their practicum experience (Steadman & Brown, 2011). As Slick (1997) finds, supervisors with a lack of knowledge over policies and procedures serve as a detriment to the student teacher, and prevent supervisors from intervening and moving student teachers from problematic placements.

Constructive Feedback: The second facet, present consistently in the literature, and in my own personal practice, is the importance of constructive feedback. Scheeler et al. (2004) finds timely, accurate feedback from the university supervisor to be one of the most effective manners for teacher candidates to improve their practice in the classroom. The student teaching experience is the ultimate exercise for teacher candidates to execute into practice what they have learned in their university education classes, and the space to reflect on how successfully (or not) they performed, so that they can grow as an educator. When teacher candidates are provided with adequate amounts of *constructive* feedback, it can act as a stimulus for reflection and promotion of critical thinking skills (Crotty & Allen, 2001; King, 2008; Napper-Owen & McCallister, 2005). It is important to distinguish that this feedback is *constructive*, as King (2008) notes that feedback can be perceived as negative, neutral, or positively, by the candidate. Additionally, feedback that is constructive is specific, so that the teacher candidate knows exactly where to improve their practice. When constructive feedback is limited, teacher candidates report this as

one of the largest negative components of their field experience (Martínez-Agudo, 2016). While the cooperating teacher has a great impact on a teacher candidate, their feedback is often reported by teacher candidates as too general (McIntyre & Killian, 1987). In contrast, the university supervisor, at its highest quality, provides targeted feedback to the teacher candidate (González-Toro et al., 2020).

Mentorship: The third facet that is located throughout the literature is the role of the university supervisor as a mentor. Multiple reports, whose authors span across the globe, emphasize the importance of mentoring in pre-service education (Darling-Hammond et al., 2017; Schleicher, 2011). Successful mentoring relationships between teacher candidates and their supervisors is often the determining factor between a successful and unsuccessful practicum experience (Ellis & Osborne, 2015; Izadinia, 2015). However, one issue that is commonplace is a lack of awareness on the part of supervisors regarding their role as a mentor (Allen & Wright, 2014). Furthermore, this aspect of mentorship emerged distinctly from the supervisor's role as an evaluator, or one providing critical/constructive feedback. The practicum experience, while viewed as the most important part of pre-service education (Cochran-Smith, 1991; Darling Hammond, 2014), is also shown to be the most stressful and emotionally taxing on teacher candidates (Izadinia, 2016). Thus, supervisors must address this, providing emotional support to candidates (Christophersen et al., 2016), as is expected by teacher candidates (Payne, 2018).

Collaboration: The fourth and final facet commonly addressed in the literature was the role of the supervisor as a collaborator between the many individuals involved in the practicum experience. Of the three roles in the student teaching triad, the university

supervisor is the individual tasked with facilitating the enactment of policies, and forging relationships between teacher candidates, cooperating teachers, K-12 students, and university professors (Donovan & Cannon, 2018). Freeman (1990) notes that the university supervisor and cooperating teacher must form a relationship that both 1) gives the teacher candidate space to explore and practice without interference and 2) gives the cooperating teacher a level of autonomy and ensures that their input is valued. From the perception of the teacher candidate, the establishment of positive relationship amongst all members of the triad is considered to be the most valued trait of the university supervisor (González-Toro et al., 2020). Furthermore, teacher candidates value when their support systems come from *both* the university supervisor and cooperating teacher, not just a single source (Ediger, 2009). Multiple studies find that the establishment of mutual trust between university supervisor and teacher candidate is important to the success of the practicum experience (Beck & Kosnik, 2002; Hobson et al., 2009; Hudson, 2016). The ability of teacher candidates to grow and exercise judgment in the classroom suffers from supervisory models in which the university supervisor lies at the center of pedagogical action (Zeichner, 2005). Thus, a university supervisor of the highest quality will successfully forge relationships amongst all parties in the practicum, and help co-construct—not dominate—the direction of the overall experience.

Step 3: Determine the Facet Levels and Generate Descriptions to Capture Variation within each Facet

With these four facets established, the next step was to define how each facet varies, from the highest levels to the lowest levels. Given that RGS scale development uses the principles of Rasch measurement, each facet must exist on a hierarchical

continuum, as one of the principles of Rasch measurement is that it is unidimensional. As shown in Table 3.1, there are three levels per facet, which I designate as *high*, *medium*, and *low*.]

Table 3.1

Facet Levels and Brief Descriptions

Facet	Level	Brief Description
Resourcefulness	High	Extremely knowledgeable about policies and procedures; initiates contact and acts as a resource
	Medium	Familiar with most policies and procedures, but not all; more passive as a resource, when called upon
	Low	Not aware of most policies and procedures; limited to no value provided as a resource
Constructive Feedback	High	Provides vast amounts of constructive, targeted feedback on lesson plans and reflections
	Medium	Provides some level of constructive feedback on lesson plans and reflections
	Low	Provides little to no constructive feedback on lesson plans and reflections
Mentorship	High	Continuously encourages teacher candidate, understands challenges of the practicum experience
	Medium	Supportive of teacher candidate
	Low	Not encouraging of teacher candidate, lacks understanding of challenges faced by candidate
Collaboration	High	Facilitates dialogue with the triad, co-constructs practicum experience
	Medium	Works alongside members of triad, can sometimes a) direct process more than others or b) not take active enough role
	Low	Rarely works with members of triad, frequently either a) directs the whole process or b) takes little to no role

Step 4: Determine the Structure of the Scenarios

One approach suggested by Ludlow, Baez-Cruz, et al. (2020), as part of the fourth step is to construct scenarios at the extreme levels, that being a scenario at the highest and lowest levels. As described previously, there are several advantages to doing this. First, it establishes clear boundaries for the scale, as all other items will be written to exist on a

continuum between these two extreme scenarios. Second, as Ludlow, Baez-Cruz, et al. (2020) describe, this exercise can be useful as serving as a "proof of concept" and ensures that the scale is rooted in theory (p. 369). Given that this scale consists of four-facets across three struts, this produces a scenario written with four facets at the highest level, and a second scenario with four facets written at a low level. In this following section, I present these two scenarios.

Scenario 1: "High" Scenario, All 4 Facets at "High" Level: Alex is extremely knowledgeable about all of the policies and procedures of the practicum experience. They provide me with vast amounts of constructive feedback on all of my lesson plans and reflections. Alex continuously encourages me in my unique role as a student teacher in my school. They facilitate dialogue between with my supervising practitioner and me to co-construct my practicum experience.

Scenario 2: "Low" Scenario, All 4 Facets at "Low" Level: Casey is aware of very few of the policies and procedures of the practicum experience. They provide me with limited to no useful feedback on my lesson plans and reflections. Casey is slightly understanding of my role as a student teacher at my school. They rarely work with my supervising practitioner and me, frequently taking an overbearing role in the process.

Step 5: Developing the Mapping Sentences and Constructing the Scenarios

Building upon the fourth step, with two extreme scenarios constructed, the fifth step was to develop mapping sentences and to construct the remaining scenarios. The sentence mapping process is informed by the work of Borg & Shye (1995) and its later application to the RGS framework, as best articulated by Ludlow, Baez-Cruz, et al. (2020). To reiterate, with sentence mapping, facets and strut levels are presented in

parentheses, and names of facets are placed onto struts within the mapping sentence. Next, informal linking phrases are placed between facets (and italicized), signifying its status as a transitionary phrase (Ludlow, Baez-Cruz, et al., 2020). Like other RGS scales, I use "unobtrusive facetization," which means facet level descriptions are *not* repeated, nor are they indicated as being a "high" or "low" strut within a scenario (Borg & Shye, 1995, p. 34). In Figure 3.1, I present my conceptualization of the mapping sentence for this scale.

Figure 3.1
Sentence Mapping, University Supervisor Quality Scale

Supervisor X is

Facet 1: Resourcefulness
(high [knowledgeable])
(medium)
(low [unaware])

of policies and procedures of the practicum experience.

They provide me with…

Facet 2: Constructive Feedback
(high [vast])
(medium)
(low [infrequent])

amounts of constructive feedback on my lesson plans and reflections.

Supervisor X is

Facet 3: Mentorship
(high [very encouraging])
(medium)
(low [slightly understanding])

of my role as a student teacher at my school.

They…

Facet 4: Collaboration
(high [continuously])
(medium)
(low [rarely])

work with my supervising practitioner and me

and

Facet 4: Collaboration continued
(high [co-construct])
(medium)
(low [take control over/absent from])

my practicum experience.

Step 6: Decide on the Response Options and Survey Instructions

The sixth step in the process was to develop survey instructions and the item response options, something incredibly important given the relative novelty of scenario-based items. By minimizing participant confusion regarding how one answers scenario items, this helps protect the instrument's validity. Like other RGS scales, this scale uses a comparative scenario response format (Ludlow et al., 2014). Unlike most traditional scale response options, where participants answer their level of agreement to a statement, these

respondents were asked to compare their experiences, perceptions, or beliefs *compared* to the scenario presented. For this scale, participants are former *teacher candidates*, who provide their perceptions of the quality of their *university supervisor*. Thus, the comparison was their perception of their university supervisor's quality compared to the quality of the supervisor presented in each scenario. My response options are as follows:

- My supervisor is much better
- My supervisor is a little better
- My supervisor and Supervisor X are about the same
- Supervisor X is a little better
- Supervisor X is much better

One feature of these response options is that they are all written with a positive slant, instead of the lower response options being "my supervisor is a little worse" and "my supervisor is much worse." The rationale for using response options with only positive wording is used as a mechanism is to fight against social desirability bias. This is because participants may feel inclined to not provide a negative response, particularly given that a supervisor-teacher candidate power dynamic may exist. This approach with solely positive language has been utilized on other RGS scales, notably Reynolds' (2020) scales, which feature university students evaluating faculty members, and where similar concerns of social desirability bias were thought to possibly pose a threat.

Step 7: Testing Congruence of Theory and Practice

The seventh and final step of RGS scale development was to administer the set of scenarios to populations to test whether the theory of the instrument is valid or not. Prior to the full administration and set of final scenario items, there were several stages where

feedback was sought from a variety of individuals. I solicited feedback from content experts in the fields of teacher education to understand their reactions to the scenario items. This included individuals in roles of in-service teachers, teacher educators, university supervisors, teacher candidate alumni, and directors of student teaching. Additionally, I sought the advice of those in the field of survey development. While not all of these individuals were teacher educators, their training in survey development and psychometrics was very helpful. Additionally, I worked with the members of my dissertation committee to review the scenarios and obtain their feedback before the final administration.

Most importantly, the scenario items were administered in a pilot-study that occurred in Fall 2018. This pilot study was conducted as a proof of concept that helped to inform revisions to items prior to the full administration that occurred in Fall 2021. The results of the pilot study are presented later in Chapter Four.

Cognitive Interviewing

One aspect of this dissertation that uniquely contributes to the emerging RGS field is the use of cognitive interviewing. Beatty (2003) provides a basic definition of cognitive interviewing as the administering of draft survey questions, while simultaneously collecting verbal information about the survey responses. The use of cognitive interviewing in the field of measurement most notably began in the 1980s, best codified in two volumes (1980 and 1993) by researchers K. Anders Ericsson and Herbert Simon. This early use of cognitive interviewing was mostly synonymous with a single technique, known as think-aloud (Ericsson & Simon, 1993). The think-aloud method centers around minimal involvement/prodding from the researcher, and to facilitate the

participant to verbalize their feelings while answering basic questions (Beatty & Willis, 2007). For example, one think-aloud question may be something along the lines of, "tell me what you are thinking…" (Beatty & Willis, 2007). While there are benefits to this form of simplistic style of interviewing (e.g. little researcher influence may help reduce bias), this can create situations where participants have no sense of direction, and ultimately cannot provide enough relevant and/or useful information (Willis, 2005).

One answer to the issues presented by the think-aloud method is to utilize probing techniques. Cognitive interviewing with probing provides the interviewer with much more control of the interview, utilizing either scripted or spontaneous responses to participant answers (Beatty & Willis, 2007). As previously mentioned, one of the benefits of the think-aloud method is its reduction in possible introduced biases, something that probing has the potential to introduce. However, the potential benefits gained from probing—either scripted or spontaneous—can outweigh these concerns over potential bias instruction, particularly if great care is taken in the pre-interview scripting process.

The analysis of data from cognitive interviews certainly is performed in a vastly different manner than the Rasch data. One technique for analyzing these data is to generate a series of qualitative codes related to various responses (Almond et al., 2009). These codes can then be transformed into descriptive statistics. However, the literature is not inherently specific about exact data analysis procedures (Ryan et al., 2012). Furthermore, in regard to sampling, there is no consensus as to what constitutes an appropriate sample size (Beatty & Willis, 2007). For a small-scale administration such as this, sample sizes will differ than that of a large-scale survey administration (Ryan et al.,

2012). Therefore, the sample size desired for the round of cognitive interviews was established between 5 and 20 participants.

Interview Protocol & Procedures

As previously described, there are multiple approaches to conducting cognitive interviews, such as think-aloud (Ericsson & Simon, 1993) as well as probing (Beatty & Willis, 2007). For the purposes of these interviews, a semi-scripted interview protocol was developed, which allowed for both think-aloud and probing techniques. The think-aloud method allows for interview participants to reflect with limited interference from the interviewer, which can help to reduce the introduction of bias. However, the inclusion of some probing mechanisms allows for the interviewer to steer conversation if the interview begins to fall off-course from the intended overall discussion.

The interview protocol included a basic introduction, in which the interviewee was welcomed to the study, explained the general format for the interview (i.e. they will see nine scenarios, they will answer the item, and then discuss their thoughts/perceptions of the item), and finally reminded that the conversation would be audio recorded (as agreed to in the consent form, signed before the interview). The interview protocol established that the conversation would take place over Zoom, in which the interviewee would see a PowerPoint presentation with each item and the item response options, one scenario at a time. Additionally, the interview protocol included basic questions for each item, such as "How did this description compare to your supervisor," "Why did you pick the response option you selected," and "Were there any words or phrases that were confusing?" These questions were repeated for each item, with the additional question added after the first, "Can you explain to me any similarities or differences you observed

while reading this scenario, compared to other scenarios?" The interview protocol allows for the interviewee to speak broadly about their thoughts and perceptions of each scenario item, as well as flexibility for the interviewer to react and follow-up on things the interviewee said over the course of the discussion.

In contrast to the online survey, in which each participant was given a completely randomized ordering of the scenarios (i.e. scenarios were presented in varying orders of item "difficulty" for each survey), the ordering of the scenarios presented on the PowerPoint was consistent across *all* interviews. Furthermore, this ordering presented varying degrees of scenario difficulty throughout the interview. This manifested itself with high, middle, and low difficulty items spread across the interview, not in a "high-to-low" or "low-to-high" ordering. The ordering of scenario items are displayed in Table 3.2.

Table 3.2

Cognitive Interviews, Scenario Ordering

Scenario Order	Scenario "Name"	Scenario Number
1	Riley	S3
2	Kyle	S7
3	Taylor	S5
4	Jordan	S2
5	Casey	S9
6	Quinn	S6
7	Alex	S1
8	Sam	S8
9	Chris	S4

CHAPTER FOUR: RESULTS

In the following chapter, I present the results from the pilot study of scenarios which served as an initial proof of concept, and whose results formed the basis for revisions for a full administration. I also present the results of the full administration. The full text for the pilot and full administration scenarios can be found in the appendix of this dissertation. As discussed in Chapter Three, these data were analyzed utilizing the principles of Rasch measurement, with each round of data analysis following the same procedures. For the purposes of this discussion, I present the results and interpretations from the pilot and full administration phases, as well as discussion from a series of 12 cognitive interviews.

Pilot Study

Overview of Responses

For the first pilot study, 216 undergraduate and graduate students from Boston College enrolled in a student teaching practicum during Fall 2018 were invited to take this survey. In collaboration with the then-named Office of Practicum Experiences and Professional Development (now the Office of Field Placement and Partnership Outreach), these individuals' names and emails were obtained through their database system. For the distribution of the survey, the survey was sent via Qualtrics, and data were collected over a six-week period from December 2018 through the end of January 2019.

Missing Data

Of the 216 individuals invited to complete the survey, 90 opened the survey link and started it, yielding an initial response rate of 41.66%. Of these 90 recorded survey respondents, 9 (10.00%) did not complete the first question, which was the statement of

consent, therefore reestablishing the sample size at 81 individuals. Of these 81, 5 (6.17%) individuals answered some of the initial demographic questions, but did not continue to the scenario items or any items beyond those as well. However, of the 76 individuals who began the scenario section of the survey, all 76 fully completed all 9 scenario items. Therefore, in respect to missing data, there are none, and so a final sample size of 76 was set.

Descriptive Statistics

Respondents

For the purposes of the pilot study, common demographic questions used in educational research (e.g. gender, race/ethnicity, etc.) were not asked of respondents to answer. However, respondents were asked to identify their program level (i.e. undergraduate or graduate), as well as the program they were enrolled in (e.g. Elementary Education, Moderate Special Needs). For all descriptive statistics that follow, I utilize the sample size of 76. This demographic information is displayed in Table 4.1.

Table 4.1

Respondent Demographic Information (Pilot)

	# of Respondents
Program Level	
Undergraduate	55
Graduate	21
Program	
Early Childhood	1
Elementary Education	38
Secondary Education	36
Severe Special Needs	1

Scenario Items

The means and standard deviations for all nine scenario items are displayed below in Table 4.2. The scenarios are ordered numerically from Scenario 1, Scenario 2…. through Scenario 9. However, this ordering is also representative of the hypothesized difficulty of the items, with Scenario 1 hypothesized to be the most difficult for respondents to provide a high score, whereas Scenario 9 was hypothesized to be the easiest for respondents to provide a high score. In order to make meaning of these means, it is important to understand how the item response categories were constructed. For each of the items, the item response categories were as follows:

5 – My supervisor is much better

4 – My supervisor is a little better

3 – My supervisor and Supervisor X are about the same

2 – Supervisor X is a little better

1 – Supervisor X is much better

From this, selecting the response option of 5—*My supervisor is much better*—translates into a higher mean score for the item. Conversely, selecting the response option of 1--Supervisor X is much better—results in a lower mean score for the item. Thus, those items with high mean scores are an indication of *easier* scenario items, whereas a lower mean item score is an indication of a *harder* scenario item. As shown in Table 4.2, scenarios were ordered according to their hypothesized order of difficulty, with the most difficult items at the top. This was also observed within the mean scores, which increased in alignment with the hypothesized structure, with Scenario 1 having the lowest item mean (2.57) and Scenario 9 having the highest item mean (4.53).

Prior to conducting the Rasch analyses, I checked for scale reliability utilizing Cronbach's Alpha. With a sample size of 76, the Cronbach's Alpha for the nine scenario items was 0.963, a strong indication of the scale items reliability.

Table 4.2

Descriptive Statistics for Scenario Items (Pilot)

	Mean	Standard Deviation
Scenario 1	2.57	1.10
Scenario 2	2.80	1.21
Scenario 3	3.05	1.21
Scenario 4	3.22	1.21
Scenario 5	3.33	1.22
Scenario 6	3.78	1.21
Scenario 7	4.29	1.07
Scenario 8	4.33	1.06
Scenario 9	4.53	0.90

Rasch Analyses

After analyzing the descriptive statistics for the scale items, the next phase was to conduct Rasch analyses. For these Rasch analyses, like most other RGS scales, I utilized the Rasch rating scale model (Andrich, 1978). The descriptive statistics displayed in Table 4.2 indicate some initial evidence of how this construct exists on a continuum, as evidenced by a corresponding item mean increase with the hypothesized increased scenario difficulty. This initial evidence is strengthened by the following Rasch analyses, producing an even more robust form of construct validation.

In the next section, I provide the most important Rasch outputs and my interpretation of those outputs. The first piece of Rasch output is the variable map, which is arguably the most important of all. The variable map provides a clear visual of the progression of difficulty of items, with the easiest items at the bottom of the map, and the

hardest items at the top. On the variable map, items should not only correspond to the hypothesized ordering, but ideally the items should have even spacing between them. Additionally, I present other Rasch output, such as the category characteristic curves, Andrich thresholds, and fit statistics. Finally, as a final measure to test unidimensionality, I conducted a principal components analysis on the item residuals.

Variable Map

As displayed in Figure 4.1, the items of the USQ scale match the hypothesized ordering from low levels of perceived supervisor quality to high levels of perceived supervisor quality. Scenario 1 (S1) is located at the top of the variable map, indicating that it was the most difficult item for participants to select response option 5, "My supervisor is much better." In contrast, Scenario 9 (S9) is located at the bottom of the variable map, which indicates it was the easiest item for respondents to select response option 5, "My supervisor is much better."

The variable map provides a clear visualization of the scale's progression from the easiest to endorse items (i.e. S9) to the hardest to endorse item (i.e. S1). In practice, this shows a progression in which the quality of the facets in each scenario are written at a higher level, and thus it becomes a harder item to endorse the top response option, "My supervisor is much better." For example, S9 was the easiest item for respondents to endorse the top option, as each of the four facets were written at the lowest level. The supervisor described in this scenario was one with the lowest levels of resourcefulness, constructive feedback, mentorship, and collaboration. Items further up the variable map with greater difficulties were evidenced by higher levels of quality within each facet. Whereas S9 was written with low levels across all four facets, the hardest item was S1,

which was written with high levels across all four facets. This supervisor was an individual with the highest levels of resourcefulness, constructive feedback, mentorship, and collaboration. The progression of the variable map is a direct result of increases in the facet levels.

As mentioned previously, item ordering is important to see on the variable map, as well as the relatively even spacing of items. While many of the items do tend to have even spacing, there are a couple of items that are more closely grouped, in addition to a couple of large gaps between items. For example, there is a tight gap between Scenarios 4 and 5, as well as Scenarios 7 and 8. While these items were in the correct order, there ideally should be a greater space between them. Additionally, the item between these two sets (i.e. Scenario 6), has much wider spacing between Scenarios 5 and 7 than the rest of the items on the scale. This presented a clear opportunity for the revision process, which sought to make Scenario 5 easier and Scenario 7 harder, thereby solving both the grouping and spacing issues simultaneously, and better capture the construct.

Additionally, the variable map displays person scores on the left side of the output. One noteworthy observation from the person scores is the grouping of five persons at the very top of the variable map. This is an indication of respondents that may have selected response option 5 for *all* nine scenarios. Likewise, there are two respondents at the very bottom of the variable map, indicating that they selected response option 1 (i.e. "Supervisor X is much better") for all or nearly all scenarios. Upon examination of the raw data, five individuals had total scores of 45 (i.e. response option 5 for all) and one individual with a total score of nine (i.e. response option 1 for all).

Fit Statistics

After examining the variable map, the next step was to examine the fit statistics. Whereas the variable map provides a visualization of the item and person distributions, the fit statistics provide quantitative results that are important to examine. As shown in Table 4.3, there are two fit statistics of interest: the information-weighted fit statistics (MSNQ-INFIT), and the unweighted fit statistic (MNSQ-OUTFIT), or the so-called Infit and Outfit MNSQ in WINSTEPS. The information-weighted statistic is an indication of rampant inconsistency of an item, meaning that many respondents provided varied responses to that item. Meanwhile, the unweighted fit statistic is an indication of extreme responses, when the observed data is vastly different than what was expected by a particular respondent. In addition to these statistics, a t statistic (ZSTD) is also provided in the output. For both mean squares, values above a range of 1.3-1.5 are thought to be problematic, whereas the criterion is above 2 for the ZSTD statistic (Linacre, 2002).

As shown in Table 4.3, three items had INFIT statistics greater than 1.3: scenarios 2, 7, and 8. At the more difficult end of the scale, scenario 2's misfit might derive from low-scoring individuals that provided unexpected high responses, whereas the misfit from scenarios 7 and 8 may be from high-scoring individuals that gave unexpected low responses. Upon examination of the most unexpected responses, this was in fact what occurred. Of the six most unexpected responses, scenarios 7 and 8 each had two unexpected responses, in which high scoring individuals gave unexpectedly low responses. On scenarios 7 and 8, two different high scoring individuals with an expected score of 5 answered 4 on each item. Additionally, one individual, who had an expected

score of 3 for both items, answered 1 on both scenarios 7 and 8. As for the OUTFIT statistic, only one item (scenario 7) had a statistic greater than 1.3, at 2.42. What this indicates is that at least one respondent provided an extremely unexpected response to this item. As just discussed, scenario 7 displayed examples of unexpected responses, observing scores of 4 and 1 when the expected scores were 5 and 3, respectively.

Figure 4.1

USQ Scale Variable Map (Pilot)

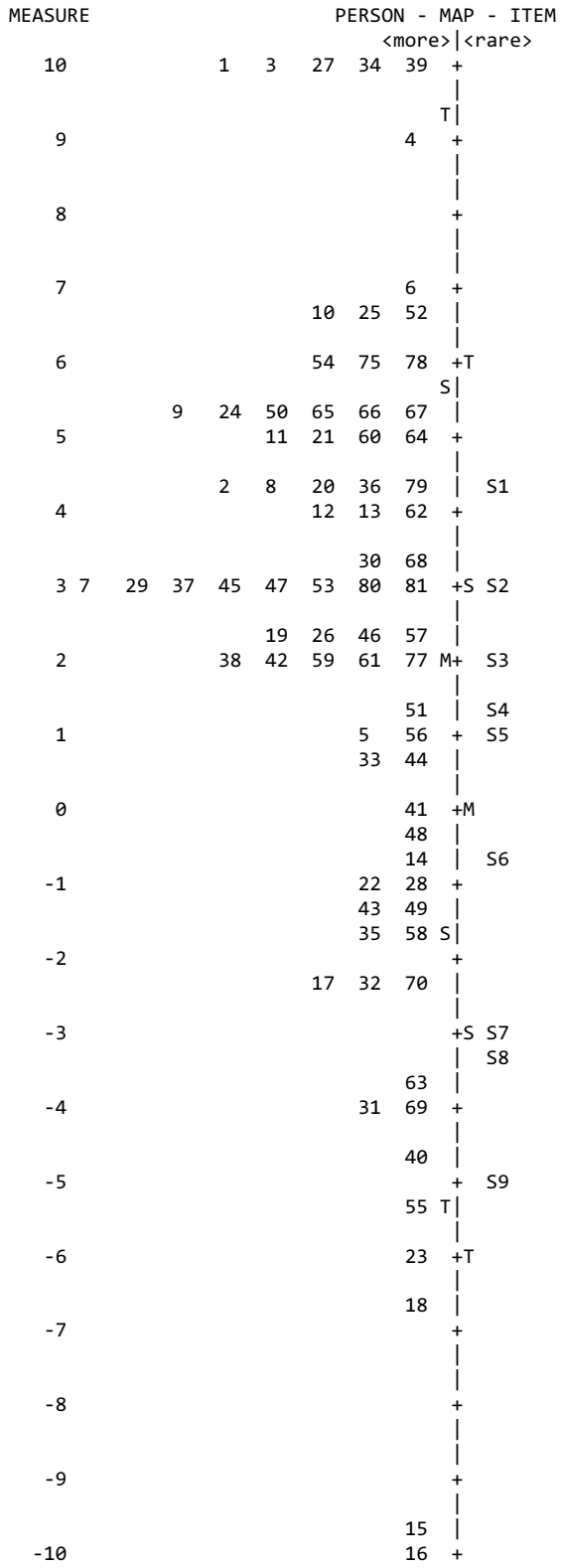


Table 4.3

USQ Scale Fit Statistics (Pilot)

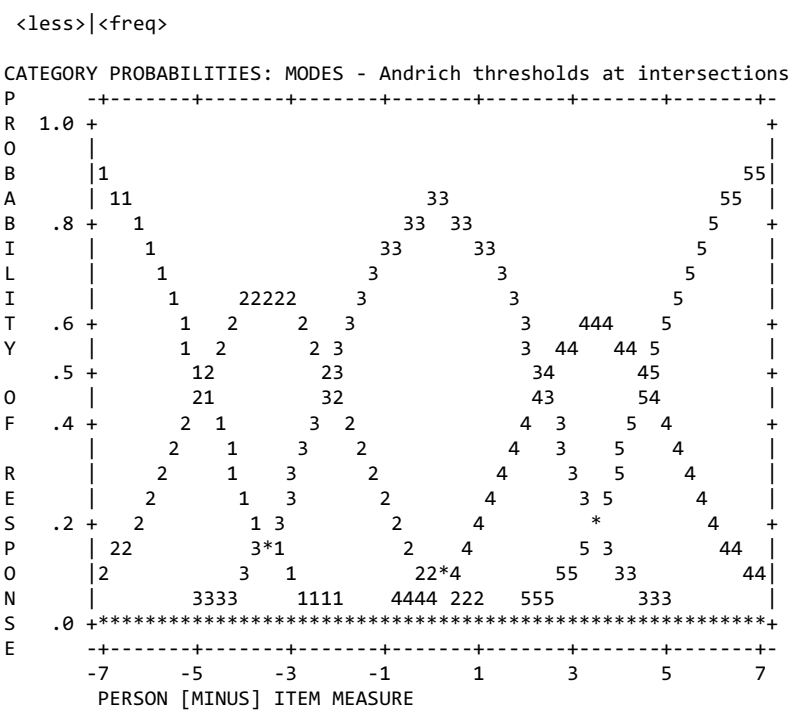
Item	Logit Estimate (S.E.)	Information-Weighted Fit Statistic		Unweighted Fit Statistic	
		MNSQ	ZSTD	MNSQ	ZSTD
Scenario 1	4.22 (.25)	.70	1.7	.60	-1.0
Scenario 2	3.16 (.24)	1.38	1.9	1.22	.8
Scenario 3	2.13 (.23)	1.01	.3	1.15	.6
Scenario 4	1.47 (.22)	.58	-2.8	.51	-.2.0
Scenario 5	1.07 (.22)	.57	-3.0	.51	-1.8
Scenario 6	-.63 (.23)	.59	-2.7	.68	-1.2
Scenario 7	-3.14 (.29)	1.54	2.2	2.42	1.9
Scenario 8	-3.40 (.29)	1.62	2.4	1.25	.6
Scenario 9	-4.89 (.34)	1.20	.8	.76	.1

Category Characteristic Curves

The category characteristic curves (CCCs) for the scale, are shown in Figure 4.2. The CCCs are a visual representation that provides additional evidence of fit. On the x-axis of the CCCs are the differences between person and item estimates, which are measured in logits. On the y-axis, are the probabilities that a respondent will select a particular response category given the difference between their person and item estimates. An ideal trajectory of the CCCs will show high probabilities of selecting response option 1 at negative person-minus-item estimates (i.e. item difficulty is greater than person difficulty), moving toward higher probabilities of selecting response option 5 at the largest positive person-minus-item estimates (i.e. person difficulty being greater than item difficulty). As shown in Figure 4.2, this pattern is generally seen.

Figure 4.2

USQ Category Characteristic Curve (Pilot)



Andrich Thresholds

The CCCs and the Andrich Thresholds shown in Table 4.4 have a direct relationship with another. The Andrich thresholds can be seen on the CCCs in Figure 4.2 at the intersection of where response categories cross one another. For example, the Andrich Threshold between response categories 1 and 2 is -4.86 (represented on the lower end), whereas between response categories 4 and 5, it is much higher at 4.58. Additional evidence of strong response categories is shown by the average estimates, which are the lowest for response option 1, and the highest for response option 5. In regard to the INFIT and OUTFIT statistics, response category 1 was the only one to eclipse the 1.3 criterion. This may be an indication that there are overall inconsistencies as well as extreme unexpected responses. As described earlier, of the most unexpected

responses, two of the top six involved cases in which one individual answered two items with a response of one, when their expected response for both was a 4. Thus, this may be the contributing factor toward this response categories' misfit. Regarding overall inconsistencies, this may be due to the overall lack of usage of this response option compared to other response categories. However, upon examination, while response category one did in fact have the lowest usage with 68 of 684 (9.94%) responses, it was relatively on par with response category two, which comprised 72 of 684 (10.53%) responses. Thus, the most likely explanation for the item's misfit lies with the one individual with the extreme responses on scenarios 7 and 8.

Table 4.4

USQ Scale Andrich Thresholds and Average Estimates (Pilot)

Response Category	Andrich Threshold	Response Frequency	INFIT	OUTFIT	Average Estimates
1 (X is much better)	NONE	68	1.45	1.58	-6.76
2 (X is a little better)	-4.86	72	.65	.59	-3.19
3 (about the same as X)	-2.15	194	.88	.92	.61
4 (a little better than X)	2.42	120	.92	1.10	3.32
5 (much better than X)	4.58	230	1.13	1.24	7.12

Principal Component Analysis

The final step in these pilot analyses was to conduct a principal components analysis (PCA), specifically on the Rasch residuals. In this instance, a residual is defined as the difference between the expected value for an individual's response (given their ability) and the actual observed response (Humphry, 2002). The main purpose of conducting a PCA on the residuals is a check on unidimensionality, and whether any other dimensions appear to exist in what is intended to be one construct. A finding of no relationship amongst the residuals is the ideal result.

The first three important statistics to examine in a PCA residual analyses are the following: the determinant of the correlation matrix, the Kaiser-Meyer-Olkin (KMO), and Bartlett's Test of Sphericity. The determinant of the correlation matrix is .102, of which a non-zero determinant is evidence of random noise. Second, as shown in Table 4.5, the KMO is .188, a small value also serving as evidence of random noise. Lastly, a non-significant Bartlett's Test would be another signal of random noise, although it is usually statistically significant (as it is in this case, p <.001).

Table 4.5

KMO and Bartlett's Test, USQ Residuals (Pilot)

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.188
Bartlett's Test of Sphericity	Approx. Chi-Square	149.022
	df	36
	Sig.	<.001

The fourth step was to examine the eigenvalues and the percentages of variance explained by those eigenvalues. As displayed in Table 4.6, the first component had an eigenvalue of 2.278, with a variance explained of 25.316%. It is important to examine the ratio of the first to second extracted components' eigenvalues (Linacre, 2022a). In this instance, the ratio of the 1st to 2nd eigenvalue (2.278/1.420 = 1.604) is *not* greater than 3/1, yet another piece of evidence of random noise (Ludlow, 2017). Additionally, the 1st component does *not* account for at least 30% of the variance (25.316%), which too is additional evidence of random noise (Ludlow, 2017).

Table 4.6

Principal Components Analyses, USQ Residuals (Pilot)

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings	
	Total	% of Variance	Cumulative %	Total	% of Variance
1	2.278	25.316	25.316	2.278	25.316
2	1.420	15.775	41.092	1.420	15.775
3	1.327	14.741	55.833	1.327	14.741
4	1.096	12.179	68.012	1.096	12.179
5	.917	10.193	78.205		
6	.747	8.305	86.510		
7	.620	6.884	93.394		
8	.491	5.456	98.850		
9	.104	1.150	100.000		

The final steps in the PCA residual analysis were to examine the scree plot and the rotated component plot. As shown in Figure 4.3, the scree plot appears to have a slight break between the first and second components, but in general, it exhibits a break-free, linear line. Second, as shown in Figure 4.4, the rotated component plot exhibits a roughly circular pattern, which has been shown to be evidence of random noise (Ludlow, 1983).

Figure 4.3

Scree Plot, USQ Residuals (Pilot)

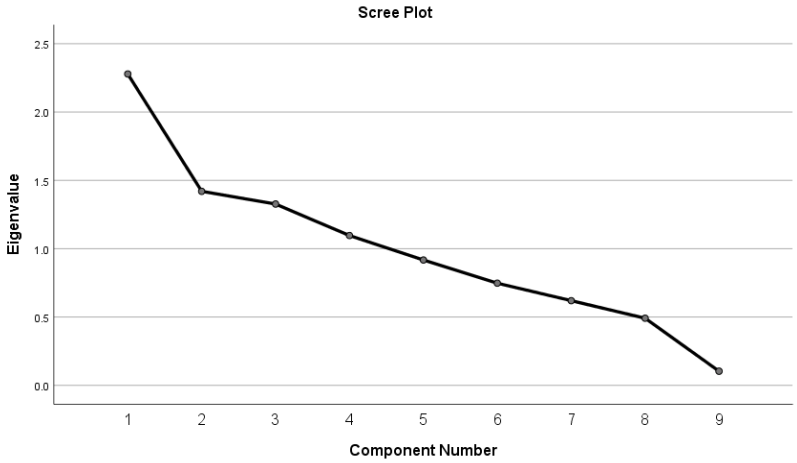
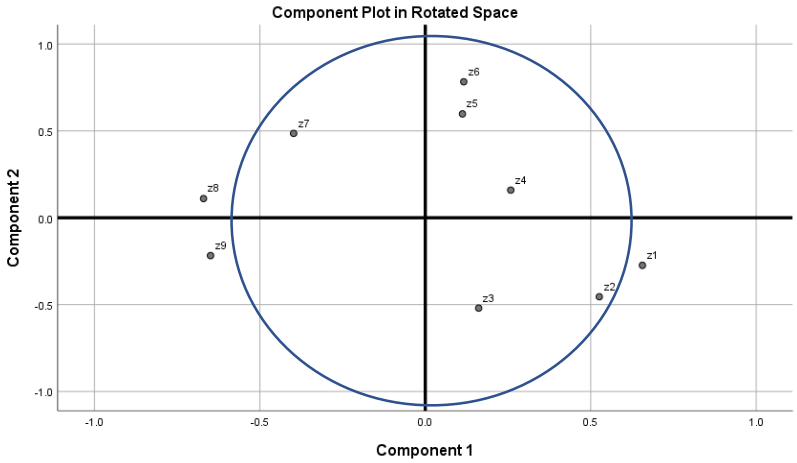


Figure 4.4

Component Plot, USQ Residuals (Pilot)



Person Level Statistics

The last piece of Rasch analyses I conducted for the pilot was to examine the person level statistics. The person level statistics offer an important compliment to many of the analyses seen prior. While reliability has been discussed earlier via Cronbach's Alpha, it is also important to examine reliability via person separation. Linacre (2022b) states that WINSTEPS produces two measures of person separation reliability, a "real" reliability and a "model" reliability. Both measures include extreme responses, but the real measure represents a lower bound of reliability, whereas the model measure is an upper bound measure. For the pilot, the person separation reliability was 0.95 (real) and 0.96 (model), a strong indication that the sample was sufficient in differentiating between high and low scorers.

Another set of person level statistics to examine are person ability estimates, as well as their measurement error. Similar to misfit analyses conducted for items, these analyses are important for identifying potential individuals whose responses either overfit or underfit the overall model. As displayed in Figure 4.5, the relationship between person ability and SEM shows a general pattern of most respondents falling between 0.60 and 1.00 for SEM. However, the plot also clearly demonstrates four outliers, two of which have high ability estimates, and two with low ability estimates. These outliers were consistent with the outliers uncovered during the previous Rasch item analyses.

Additionally, another source of important person level statistics are the person fit statistics. As shown in Figure 4.6, there are four individuals (5.26%) that appear to be outliers compared to the rest of the sample. This is another indication of the patterns observed in the previous analyses.

Figure 4.5.

Person SEM vs. Person Ability Estimates (Pilot)

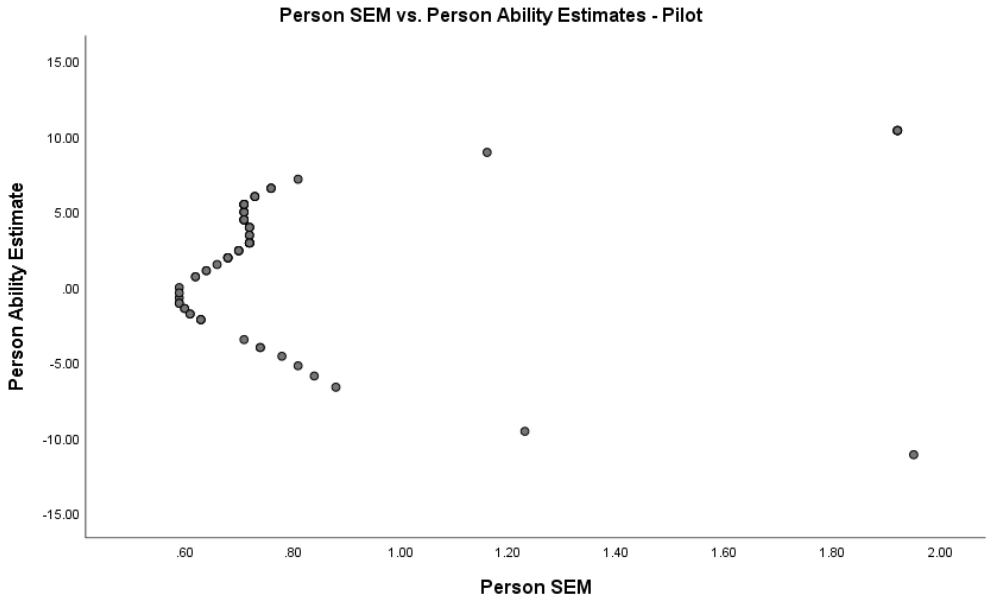
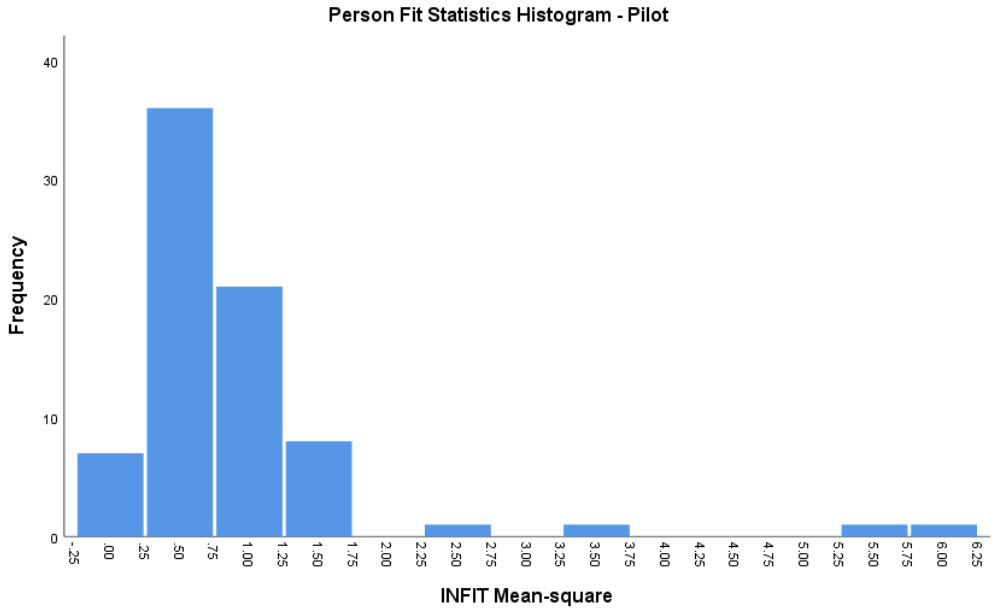


Figure 4.6

Person Fit Statistics Histogram, (Pilot)



Item Revisions

While the analyses described above demonstrated the "proof of concept," they also presented clear opportunities for improvement. Displayed in Table 4.7 are the item revisions that took place between the pilot and full administration studies. Listed in Table 4.8 are the pilot scenario items, full administration items, and the rationale for the revision. Additionally, item revisions made are bolded in the full administration column. The ordering of items is based on their hypothesized degree of item difficulty, from highest to lowest.

Every item had revisions made from the pilot to the full administration. The most substantial revision was the addition of language at the start of every item's second sentence. This language addresses the act of being a resource, which was added to better capture the facet of "resourcefulness." Previously, this first facet had been labeled as "knowledgeable." However, examination of the Rasch results and discussion of them with content experts spurred discussion over how one could be a *knowledgeable* supervisor, but fail to *dispense* that knowledge onto their teacher candidates. Thus, the true measure of a supervisor's quality in this area is better captured by addressing both their knowledge of policies and procedures *and* the frequency in which they act as a resource.

Additionally, as evidenced by the variable map in Figure 4.1, 23 of 76 (30.26%) individuals scored *above* the most difficult scenario, S1. This was a clear indication that this item needed revision to increase its difficulty, as to better capture the overall construct. The addition of language mentioned in the previous paragraph was part of the solution to increase this item's difficulty, as well as changing the phrase "continuously

encourages me" to "always encourages me." However, in order to maintain the ideal spacing between this item and the following two that was seen in the pilot, the language of these two scenarios were also modified to be slightly more difficult.

The greatest areas of concern were the middle items on the variable map, S4, S5, and S6. S4 was shown to be too similar in difficulty to S3 and S5, which indicated a need for revisions to make both S4 and S5 easier to endorse the top response selection. Additionally, of these two, S5 required even further modification to differentiate itself from the newly revised easier S4. As shown in the variable map in Figure 4.1, S6 has the largest gaps between its neighboring items, S5 and S7. The modifications made to S5 were intended to bring this item closer in difficulty to S6, while language was added to S7 to make it slightly more difficult. This language addition to S7 also provided a benefit of differentiating it more from S8, which it was too similar to.

Table 4.7

USQ Item Revisions Based on Pilot Study

Pilot Item	Full Administration Item	Rationale for Item Revisions
Alex is extremely knowledgeable about all of the policies and procedures of the practicum experience. They provide me with vast amounts of constructive feedback on all of my lesson plans and reflections. Alex continuously encourages me in my unique role as a student teacher in my school. They facilitate dialogue between with my supervising practitioner and me to co-construct my practicum experience.	Alex is extremely knowledgeable about all the policies and procedures of the practicum experience. **They continuously act as a vital resource**, **frequently providing** me with vast amounts of constructive feedback on all my lesson plans and reflections. Alex **always** encourages me in my unique role as a student teacher in my school. They facilitate dialogue between my supervising practitioner and me to co-construct my practicum experience.	Due to 23 of 76 (30.26%) of person's scoring *above* this item, it was determined that this item needed to be revised to be more difficult. There were three revisions made to this item to increase its difficulty. The first was the addition of language about being a resource, and done at the highest level (i.e. "continuously," "vital"). The second revision was the modification to "frequently" providing feedback over just providing. Lastly, the mentorship facet was revised by exchanging the word "continuously" for "always."
Jordan is well-informed about the policies and procedures of the practicum experience. They provide me with extensive constructive feedback on my lesson plans and reflections. Jordan encourages me in my unique role as a student teacher. They partner closely with my supervising practitioner and me throughout my practicum experience.	Jordan is well-informed about the policies and procedures of the practicum experience. **They frequently act as an important resource**, **providing** me with extensive constructive feedback on my lesson plans and reflections. Jordan encourages me in my unique role as a student teacher. They partner closely with my supervising practitioner and me throughout my practicum experience.	This item had only one revision, which was the addition of the resourcefulness language. This language was written to be at a high level, albeit slightly below S1. The modifiers used for this act was "frequently" and "important resource."
Riley is informed about the policies and procedures of the practicum experience.	Riley is **well-informed** about the policies and procedures of the practicum experience.	There were three revisions to this item. The first revision was to the resourcefulness facet, which made it slightly

They provide me with substantial constructive feedback on my lesson plans and reflections. Riley is considerate of my unique role as a student teacher at my school. They collaborate with my supervising practitioner and me during my practicum experience.	**They often act as a valuable resource**, **providing** me with substantial constructive feedback on my lesson plans and reflections. Riley is **very** considerate of my unique role as a student teacher at my school. They collaborate with my supervising practitioner and me during my practicum experience.	more difficult, modifying from "informed" to "well-informed." Additionally, the new language about acting as a resource was added. This was also written at a high level, but slightly below S1 and S2, as they "often act as a valuable resource." Lastly, the mentorship facet was made slightly more difficult, adding the modifier "very" to the word considerate.
Chris is familiar with most of the policies and procedures of the practicum experience. They provide me with considerable amounts of valuable feedback on my lesson plans and reflections. Chris is very supportive of my role as a student teacher at my school. They work alongside my supervising practitioner and me during my practicum experience.	Chris is **informed of most** of the policies and procedures of the practicum experience. **They act as an important resource**, **providing** me with considerable amounts of valuable feedback on my lesson plans and reflections. Chris is supportive of my role as a student teacher at my school. They work alongside my supervising practitioner and me during my practicum experience.	This item had two revisions. The first revision was to the resourcefulness facet, which slightly increased the difficulty, replacing the word "familiar" with "informed of." Additionally, the language of acting as a resource was added, intended to be at a middle level. The level of "act" is not specified either positively or negatively, nor is the level of "providing" specified. These were intended to indicate a moderate level.
Taylor is familiar with policies and procedures of the practicum experience. They provide me with helpful feedback on my lesson plans and reflections. Taylor supports me in my role as a student teacher at my school. They work with my supervising practitioner and me during my practicum experience.	Taylor is **mostly familiar** with the policies and procedures of the practicum experience. They **can be a useful resource**, **providing** me with helpful feedback on my lesson plans and reflections. Taylor supports me in my role as a student teacher at my school. They **generally** work with my supervising practitioner and me during my practicum experience.	This item had three revisions. This scenario was too close in difficulty to S4, and thus revisions were made to make it an easier item. The first revision was to the resourcefulness facet, which modified "familiar" to the slightly worse "mostly familiar." Additionally, the new resourcefulness language was added, with the intent of being slightly below average, evidenced by "can be a useful resource." Lastly, the word "generally" was added in the collaboration

		facet, to make it slightly easier.
Quinn is aware of most policies and procedures of the practicum experience. They provide me with some helpful feedback on my lesson plans and reflections. Quinn is understanding of my role as a student teacher at my school. They work with my supervising practitioner and me, but sometimes they seem to direct the process.	Quinn is **mostly aware** of the policies and procedures of the practicum experience. **They sometimes act as a resource**, **providing** me with some helpful feedback on my lesson plans and reflections. Quinn is **mostly** understanding of my role as a student teacher at my school. They work with my supervising practitioner and me, but sometimes they seem to direct the process **or provide little input**.	Given the large space between S5 and S6 on the pilot, S5 was revised to be an easier item. However, it was important to revise S6 only slightly, as to maintain its distance from this newly revised item. Thus, there were three item revisions. The first was to the resourcefulness facet, which modified the level to "mostly aware" from "aware," similar to the previous scenario. Second, the addition of the resourcefulness language was added, intended to be at a slightly below average level, as evidenced by the phrase "sometimes act." Lastly, language was added at the end to clarify the collaboration facet. The pilot item only described a supervisor that sometimes would "direct the process." What this is an indication of is a slightly less than ideal enactment of collaboration. However, it was important to include the counterpart to "direct the process" with "provide little input," as that is also a slightly less than ideal enactment of the facet.
Kyle is unaware of some policies and procedures of the practicum experience. They provide me with selected amounts of useful feedback on my lesson plans and reflections. Kyle is somewhat	Kyle is **aware** of some, **but not all**, policies and procedures of the practicum experience. **They sporadically act as a resource**, **providing** me with selected amounts of useful feedback on my lesson plans and	This item had four revisions. The first revision was to change the word "unaware" to the positive "aware," but modified by the phrase "but not all" after. This was done to be consistent with other items, in which the verbs/actions are positive,

understanding of my role as student teacher at my school. They sometimes work with my supervising practitioner and me, but they also can take too commanding a role in the process.	reflections. Kyle is **normally** understanding of my role as a student teacher at my school. They sometimes work with my supervising practitioner and me, but they **either** take too commanding a role**, or an insufficient amount,** in the process.	and modified by adjectives/adverbs to distinguish quality and/or frequency. Second, the act of being a resource language was added, which was intended to be at a lower level, evidenced by the word "sporadically." Third, the mentorship facet was made to be slightly better, opting for the word "normally" over "somewhat" understanding. Lastly, language was added (similar to the previous item) to better capture the collaboration facet. This added the language "either" prior to the phrase "too commanding a role," and added "or an insufficient amount" to mirror it.
Sam is aware of few policies and procedures of the practicum experience. They provide me with limited amounts of useful feedback on my lesson plans and reflections. Sam is mostly understanding of my role as student teacher at my school. They infrequently work with my supervising practitioner and me, often taking too dominant a role in the process.	Sam is aware of few policies and procedures of the practicum experience. **They infrequently act as a resource**, **providing** me with limited amounts of useful feedback on my lesson plans and reflections. Sam is **somewhat** understanding of my role as a student teacher at my school. They infrequently work with my supervising practitioner and me, **either** taking too dominant a role**, or are too removed from** the process.	This item had three revisions. The first revision was the addition of the resource language, which was written at a low level, shown by "infrequently" act. The mentorship facet was written to be slightly easier, modifying the level of understanding from "mostly" to "somewhat." Finally, the collaboration facet was further clarified by adding "either" and the phrase "or are too removed from" to counter "taking too dominant a role."
Casey is aware of very few of the policies and procedures of the practicum experience. They provide me with limited to no useful feedback on my lesson	Casey is aware of very few policies and procedures of the practicum experience. **They rarely act as a resource**, **providing** me with limited to no useful feedback on my lesson	The final item had three main revisions. The first revision was the addition of the resource language, written at the lowest level, evidenced by "rarely" act. The mentorship facet was

plans and reflections. Casey is slightly understanding of my role as a student teacher at my school. They rarely work with my supervising practitioner and me, frequently taking an overbearing role in the process.	plans and reflections. Casey is **not** understanding and **can be discouraging in** my role as a student teacher at my school. They rarely work with my supervising practitioner and me, **either** frequently taking an overbearing role**, or are completely absent from** the process.	also written to be easier, with changes from "slightly" to "not" understanding, and adding the phrase "can be discouraging." Finally, the collaboration facet added the word "either" and the phrase "or are completely absent from" to parallel the phrase "frequently taking an overbearing role."

Full Administration

Overview of Responses

For the full administration study, 364 undergraduate and graduate student alumni of Boston College who were enrolled in a student teaching practicum during the graduating classes of 2019, 2020, and 2021 were invited to take this survey. In collaboration with the Office of Field Placement and Partnership Outreach, these individuals' names and emails were obtained through their database system. The survey was sent via Qualtrics, and data were collected over a four-week period from early November 2021 through the beginning of December 2021. In the survey invitation, a universal survey-link was also provided for participants to forward to other potential eligible participants.

Missing Data

Of the 364 individuals invited to complete the survey, 68 opened the survey link and signed the consent form, yielding an initial response rate of 18.68%. Additionally, 6 individuals opened the survey via the universal-link provided in the initial email invite, and signed the consent form. Of these 74 respondents, 13 (17.57%) electronically signed the consent form, but did not continue to the scenario items or any items beyond those as well, thus establishing a new sample size of 61. Of the 61 individuals who began the scenario section of the survey, 60 (98.36%) fully completed all 9 scenario items, and 1 individual (1.64%) answered 8 of 9 scenarios. In respect to this minor piece of missing data, this individual's responses are included in all Rasch and descriptive statistical reporting, with the exception of calculating the scale's Cronbach's Alpha reliability, in which a list-wise deletion removed the case from calculation.

Descriptive Statistics

Respondents

The full administration, like the pilot study, did not ask respondents to provide demographic questions such as gender, race/ethnicity, or socio-economic status. However, respondents were asked to identify the year they completed their last student teaching, program level (i.e. undergraduate or graduate), and degree program (e.g. Elementary Education, Moderate Special Needs). Additionally, participants were asked whether they are currently still employed as a teacher. For all descriptive statistics that follow, I utilize the sample size of 61. This demographic information is displayed in Table 4.8.

Table 4.8

Respondent Demographic Information (Full Administration)

	# of Respondents
Practicum Year	
2018	3
2019	16
2020	29
2021	13
Program Level	
Undergraduate	37
Graduate	24
Program	
Early Childhood	1
Elementary Education	25
Moderate Special Needs	9
Reading	2
Secondary Education	20
Severe Special Needs	4
Currently Teaching	
Yes	50
No	11

Scenario Items

The means and standard deviations for all nine scenario items are displayed below in Table 4.9. The scenarios are ordered numerically from Scenario 1, Scenario 2…. through Scenario 9. As with the pilot study, the item ordering is listed in decreasing degrees of their hypothesized difficulties. The item difficulties, which manifested themselves in the hypothesized order in the pilot study, also were seen in the full administration. Scenario 1 had the lowest item mean (2.13) and Scenario 9 had the high item mean (4.43).

Table 4.9

Descriptive Statistics for Scenario Items (Full Administration)

	N	Mean	Standard Deviation
Scenario 1	61	2.13	.87
Scenario 2	60	2.28	.94
Scenario 3	61	2.54	.81
Scenario 4	61	2.92	.95
Scenario 5	61	3.11	1.03
Scenario 6	61	3.59	1.01
Scenario 7	61	4.02	.96
Scenario 8	61	4.23	.80
Scenario 9	61	4.43	.76

Like the analyses conducted for the pilot, I checked the scale's reliability prior to conducting the larger Rasch analyses, via Cronbach's Alpha. As previously mentioned, one individual did not provide a response to one scenario item, and thus the calculation for scale reliability utilized a sample size of 60. For the full administration, the Cronbach's Alpha for the nine scenario items was 0.925, an indication of high reliability for the scale.

Rasch Analyses

Like the pilot study, I utilized the Rasch rating scale model to analyze the results of the scale (Andrich, 1978). The results presented below reflect the same procedures as the pilot study, which include examination of the variable map, fit statistics, person-level statistics, and a principal components analysis of the Rasch residuals.

Variable Map

As displayed in Figure 4.7, the items of the USQ scale match the hypothesized ordering from low levels of perceived supervisor quality to high levels of perceived supervisor quality. Scenario 1 (S1) is located at the top of the variable map as it was in the pilot. However, one of the goals of the item revisions was to increase the difficulty of this item, while maintaining it as the most difficult. In the pilot study, 23 of 76 (30.27%) individuals had total scores above this highest scenario, whereas only 5 of 61 (8.20%) were in the full administration. This is a marked improvement, which indicates that the added language did contribute to increasing the item's difficulty. In contrast, Scenario 9 (S9) is located at the bottom of the variable map, which indicates it was the easiest item for respondents to select response option 5, "My supervisor is much better."

Fit Statistics

After examining the variable map, the next step in my Rasch analyses was to examine the fit statistics.

As shown in Table 4.10, 2 items had INFIT statistics greater than 1.3: Scenarios 1 and 2. At the more difficult end of the scale, scenarios 1 and 2's misfit might derive from low-scoring individuals that provided unexpected high responses. Upon examination, the individual with the greatest misfit from the sample was in fact a low-scoring individual

(21/45 total score, logit measure -1.98) individual who provided unexpectedly high responses (both 4s) to scenarios 1 and 2, when their expected responses were 1.24 and 1.35, respectively. As for the OUTFIT statistic, scenarios 1 and 2 had a value greater than 1.3. This result can be attributed to the aforementioned individual who provided an extremely unexpected response to these items. Nevertheless, given the overall stability of the scale, no individuals were excluded from the final analyses.

Figure 4.7
USQ Scale Variable Map (Full Administration)

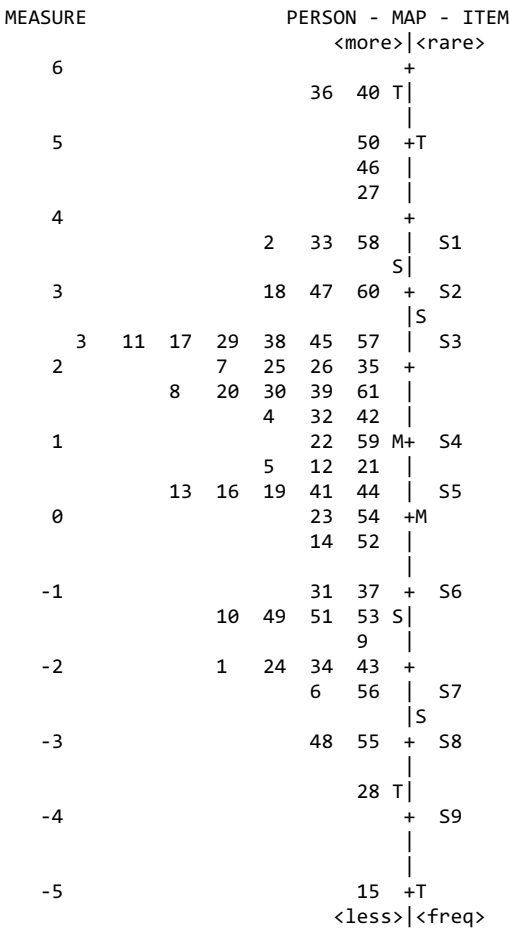


Table 4.10

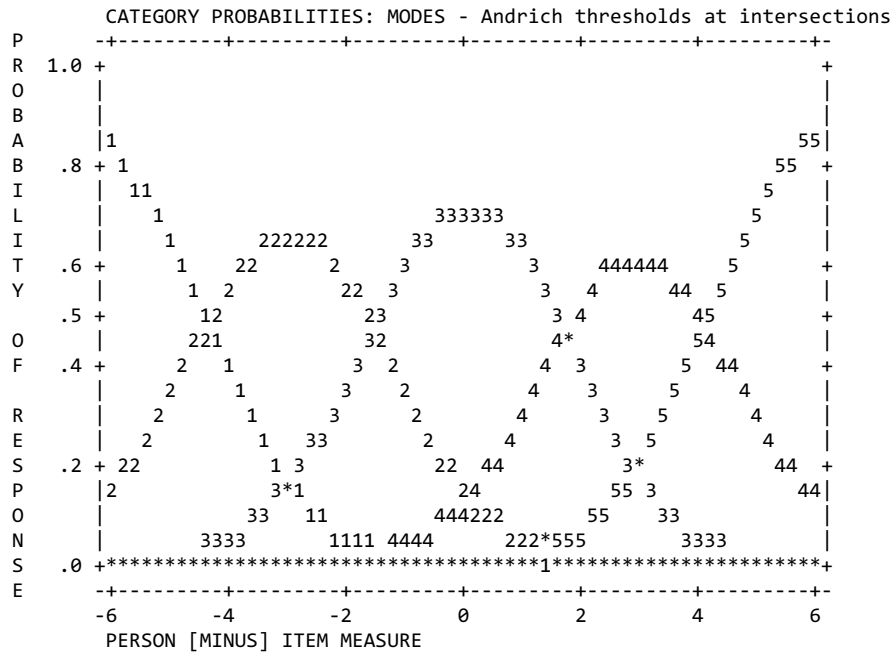
USQ Scale Fit Statistics (Full Administration)

Item	Logit Estimate (S.E.)	Information-Weighted Fit Statistic		Unweighted Fit Statistic	
		MNSQ	ZSTD	MNSQ	ZSTD
Scenario 1	3.57 (.24)	1.46	2.3	1.60	2.2
Scenario 2	3.05 (.23)	1.45	2.2	1.47	2.0
Scenario 3	2.22 (.23)	.49	-3.5	.48	-3.2
Scenario 4	1.04 (.22)	.72	-1.6	.70	-1.8
Scenario 5	.45 (.22)	.73	-1.6	.71	-1.7
Scenario 6	-.98 (.22)	1.29	1.6	1.20	1.0
Scenario 7	-2.33 (.24)	1.03	.2	1.05	.3
Scenario 8	-3.10 (.25)	.69	-1.7	.59	-1.2
Scenario 9	-3.91 (.27)	1.08	.5	1.02	.2

Category Characteristic Curves

The next source of Rasch output I examined was the category characteristic curves (CCCs) for the scale, as shown in Figure 4.8. The points of intersection between curves are the Andrich thresholds (shown in Table 4.12), which are the probabilities of moving between response categories. Specifically, it is a logit estimate where there is a .5 probability of a respondent moving from one response category to the next. For example, the Andrich threshold for response category 3 (i.e. *My supervisor and X are about the same*) is a logit estimate where a respondent has a predicted .5 probability of selecting the third response category or the second response category (i.e. *Supervisor X is better)*. Figure 4.6 demonstrates that all five response categories were chosen throughout the scale. Additionally, there were points along the continuum where each response category had a probability of being selected above .5, an indication that each response category functioned as intended.

Figure 4.8
USQ Category Characteristic Curves (Full Administration)



Andrich Thresholds

The Andrich threshold can be seen on the CCCs in Figure 4.6 at the intersection of when response probabilities cross one another. For example, the Andrich Threshold between response categories 1 and 2 is -4.35 (represented on the lower end), whereas between response categories 4 and 5, it is much higher at 4.13. Additional evidence of strong response categories is shown by the average estimates, which are the lowest for response option 1, and the highest for response option 5. In regard to the INFIT and OUTFIT statistics, the response category 1 was the only one to eclipse the 1.3 criteria, with mean squares of 1.59 and 1.54, respectively. This may be an indication that there are overall inconsistencies as well as extreme unexpected responses when a response of 1 was given. For example, respondents who typically provided high response scores may have responded with this low response category, resulting in item misfit. However, upon

examination, the first use of response category one among top scorers was by an individual in the 65th percentile, suggesting the reason for response category one's misfit was due to unexpected responses of 1 when a score of 2 or greater was expected.

Table 4.11

USQ Scale Andrich Thresholds and Average Estimates (Full Administration)

Response Category	Andrich Threshold	Response Frequency	INFIT	OUTFIT	Average Estimates
1 (X is much better)	NONE	46	1.59	1.54	-4.68
2 (X is a little better)	-4.35	104	.83	.74	-2.31
3 (about the same as X)	-1.56	171	.89	1.04	.16
4 (a little better than X)	1.78	120	1.08	.96	2.53
5 (much better than X)	4.13	107	.80	.86	5.55

Principal Component Analysis

The final step was to conduct a principal components analysis (PCA) on the Rasch residuals. The determinant of the correlation matrix is .007, evidence of random noise. Second, as shown in Table 4.12, the KMO is .129, a small value also serving as evidence of random noise. Lastly, a non-significant Bartlett's Test would be another signal of random noise, although it is usually statistically significant (as it is in this case, p <.001).

Table 4.12

KMO and Bartlett's Test, USQ Residuals (Full Administration)

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.129
Bartlett's Test of Sphericity	Approx. Chi-Square	279.242
	df	36
	Sig.	<.001

As displayed in Table 4.13, the first component extracted had an eigenvalue of 3.135, with a variance explained of 34.831%. The ratio of the 1st to 2nd eigenvalue (3.135/1.506) is *not* greater than 3/1, which is evidence of random noise (Linacre, 2022a; Ludlow, 2017). The 1st component does account for at least 30% of the variance, but this is situated amongst a litany of other indications of non-random noise, and thus is not cause for concern (Ludlow, 2017).

Table 4.13

Principal Components Analyses, USQ Residuals (Full Administration)

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings	
	Total	% of Variance	Cumulative %	Total	% of Variance
1	3.135	34.831	34.831	3.135	34.831
2	1.506	16.728	51.559	1.506	16.728
3	1.099	12.213	63.773	1.009	12.213
4	.900	9.997	73.769		
5	.842	9.356	83.125		
6	.744	8.271	91.397		
7	.501	5.566	96.962		
8	.255	2.831	99.794		
9	.019	.206	100.000		

As shown in Figure 4.9, the scree plot appears to have a slight break between the first and second components, but in general, it exhibits a break-free, linear line. Second, as shown in Figure 4.10, the rotated component plot exhibits a roughly circular pattern, which has been shown to be evidence of random noise (Ludlow, 1983).

Figure 4.9

Scree Plot, USQ Residuals (Full Administration)

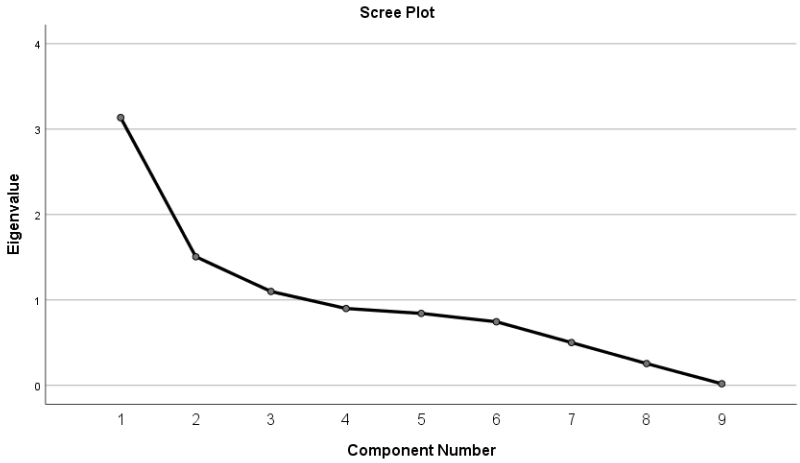
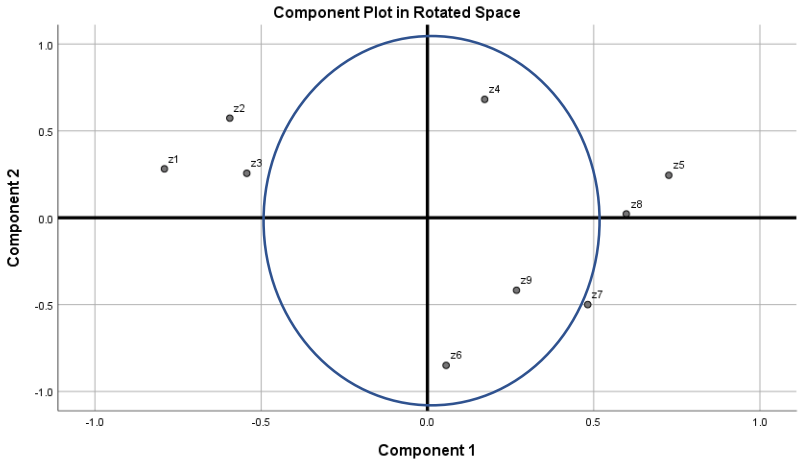


Figure 4.10

Component Plot, USQ Residuals (Full Administration)



Person Level Statistics

The final Rasch analysis conducted was to examine person level statistics. As described earlier, these statistics compliment much of what was discussed during the previous item Rasch analyses. For the final administration, the person separation reliability was 0.90 (real) and 0.93 (model), another indication that this sample was sufficient in differentiating between high and low scorers.

Additionally, I examine person ability estimates and their measurement error, the purpose of which is to identify responses that may overfit or underfit the overall model. As displayed in Figure 4.11, the relationship between person ability and SEM shows a general pattern of respondents falling between 0.50 and 0.75 for SEM. For the pilot administration, this chart (Figure 4.5) clearly demonstrates four outliers, while this plot does not appear to have any glaring outliers. This is a clear improvement from the pilot, and directly relates to the improvement that was seen regarding item spacing, best displayed by the full administration's variable map (Figure 4.7).

Lastly, I examine the distribution of the person fit statistics. As shown in Figure 4.12, there are four individuals (6.67%) that appear to be outliers compared to the rest of the sample, with one very extreme outlier. This finding is consistent with the identification of outliers observed in the previous item Rasch analyses.

Figure 4.11.

Person SEM vs. Person Ability Estimates (Full Administration)

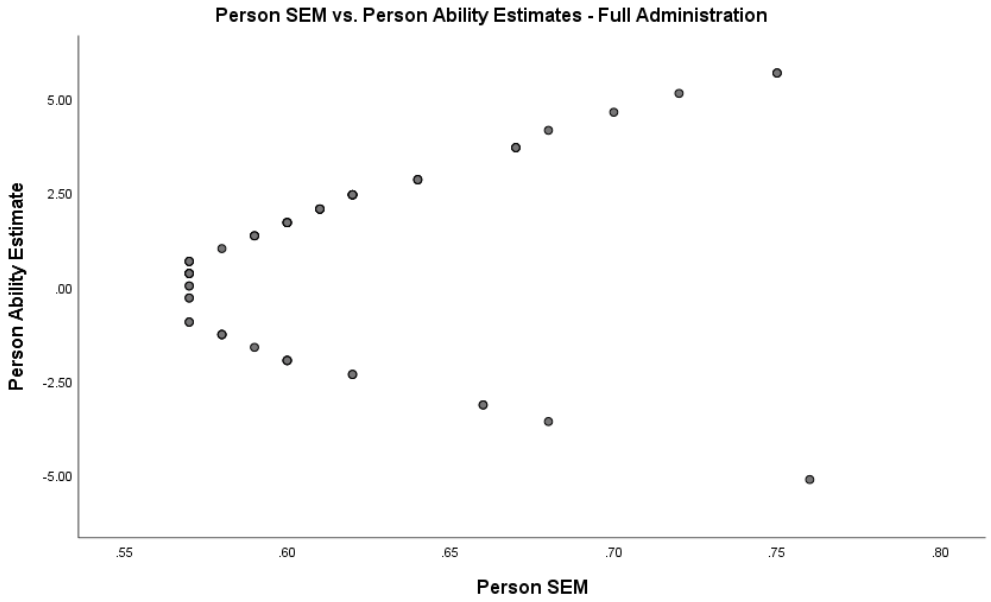
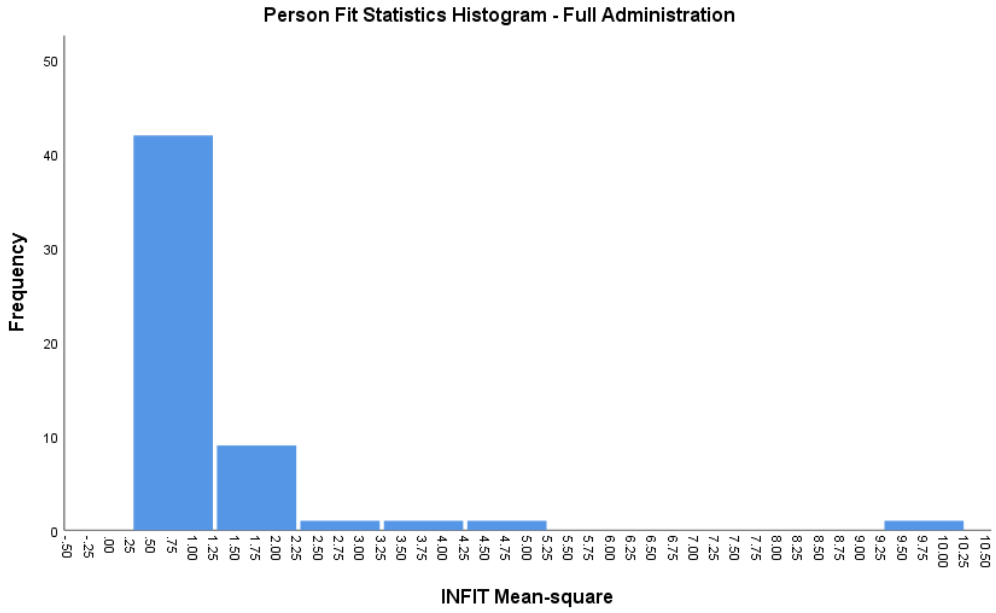


Figure 4.12

Person Fit Statistics Histogram, (Full Administration)



Comparison of Pilot Study Versus Full Administration

There were several areas of improvement in the full administration from the pilot study. The variable map from the pilot study demonstrated some issues that warranted changes in hopes that the full administration would resolve those problems. First, there were many individuals above the hardest scenario (S1), a sign that the item needed to be made more difficult, as it was not capturing this top end of the construct. The results from the full administration displayed that S1 was a much more difficult item. Second, there were issues in spacing for several of the items in the pilot study. The spacings between S3, S4, and S5 were very close together, a possible sign of redundancy. For the full administration, changes in language were made to better differentiate them from one another. As seen on the variable map (Figure 4.7), the spacing for these items was much more spread out for the full administration. However, the largest spacing issue from the pilot study was the large gap between S5 and S6, and S6 and S7. Additionally, the spacing between S7 and S8 was also extremely narrow. Changes to language in all of these scenarios were made, specifically designed to make S5 an easier item, and S7 a more difficult item, resulting in smaller gaps between them and S6. The full administration variable map did in fact demonstrate that these language changes made much more even spacing between all of these items. Lastly, the easiest item of the scale, S9, had many individuals below it on the pilot study, suggesting that it should be made even easier, to better capture the bottom end of the construct. Thus, language was changed to make S9 an easier item to endorse, which was a supervisor of even worse quality. The full administration showed that S9 was in fact a much easier item, with only one individual falling beneath it.

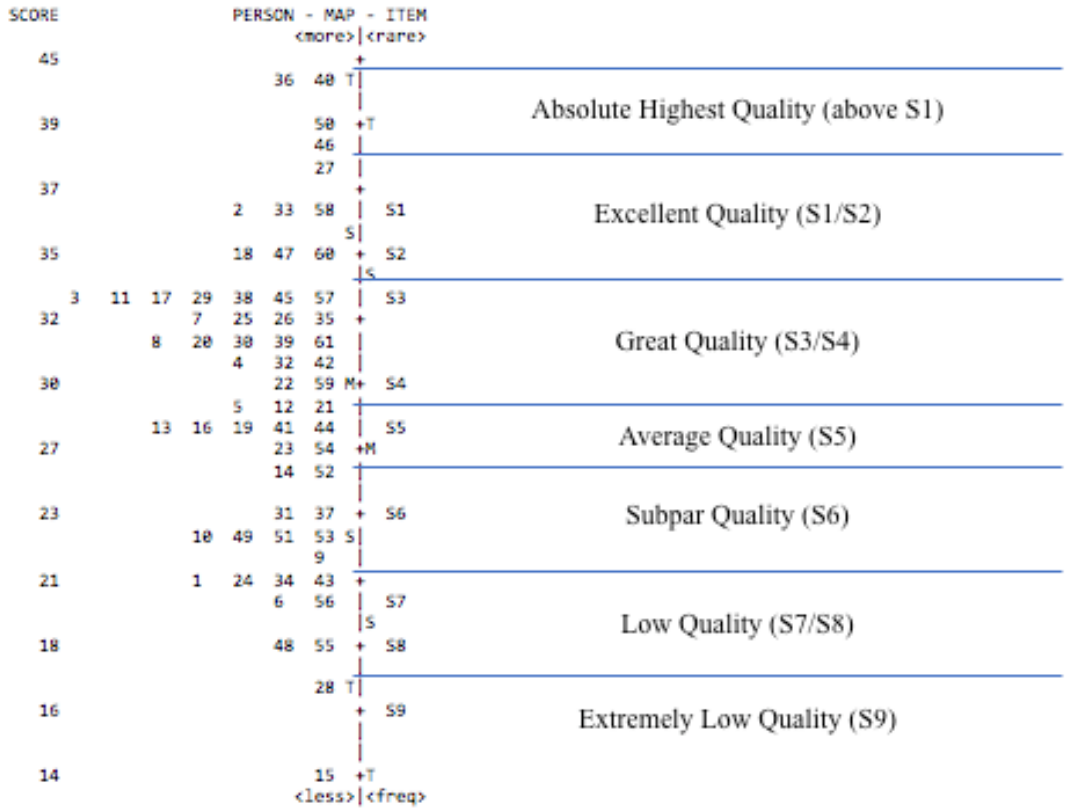
Another area of improvement between the pilot study and the full administration was in regard to the scenario fit statistics. In the pilot study, three of the items demonstrated a level of misfit, whereas on the full administration only two of the items demonstrated misfit. While ideally there would be no items with levels of misfit, this demonstrated an improvement to the scale's psychometric properties. Additionally, there were areas where no change occurred between the pilot study and full administration. One example of this is evidenced by the Andrich thresholds, in which response category one showed signs of misfit in both studies. There was one metric however that became worse in the full administration, which was the person-level misfit. In the pilot study, the percentage of misfits was 13.16%, whereas it was 14.75% in the full administration, evidenced by mean-squares above 1.40. However, the percentage of mean-squares less than 0.60 (a sign of consistency in the data) increased from 52.63% to 62.30%, which is a welcomed improvement for a Guttman scale.

Interpretations and Implications of Final Administration Results

One of the greatest strengths of scales constructed using the RGS framework is that total scores are easily accompanied with a high level of interpretability. This is because the principles of Rasch measurement, when adhered to, result in both individuals and items existing alongside one another on a single continuum. As shown in Figure 4.13, an individual's total score places them in proximity to particular items, which can be easily interpreted as providing an overall description of the supervisor's quality.

Figure 4.13

USQ Scale Variable Map, with Scores & Categories (Full Administration)



For ease of interpretability, the variable map shown in Figure 4.7 is different than previous variable maps. The far-left category was changed to "score," rather than the more technical Logit measure. Additionally, I assigned a categorical description for particular ranges of scores, from "absolute highest quality" to "extremely low quality." The score ranges from 14 (frequent selection of response category one, "X is much better than my supervisor) to 45, (maximum selection of response category five, "My supervisor is much better than X").

At the bottom of the variable map in Figure 4.9, those individuals who scored in the range of 14 to 18 indicate that their university supervisor's quality was much lower than what was presented in the nine scenarios. For those scoring in this range, individuals

describe their supervisor as about the same or slightly below someone that a) is aware of very few policies/procedures, b) rarely provides useful feedback, c) is not very understanding of the TC role, and d) rarely works with the TC and SP, by being either too overbearing or completely absent from the process. This represents a supervisor that is of "extremely low quality," of which very few scores emerged in this range.

Above this zone are individuals whose total score ranged from 19 to 23 and who indicate that their university supervisor's quality was lower than most of the other scenarios, although slightly above the worst scenario (i.e. S9). Individuals who scored in this range describe their supervisor as about the same as someone who a) is aware of few policies/procedures, b) provides limited constructive feedback, c) is only somewhat understanding of the TC role, and d) infrequently works with the TC and SP, either by being too commanding or too absent from the process. This range is an indication of a supervisor that is of "low quality."

Next, the area above "low quality" represents individuals who score in the range of 24 to 28, which is below the highest scenarios, but slightly above the lowest scenarios. Individuals in this range describe their supervisor as about the same as someone who is a) mostly aware of policies/procedures, b) sometimes provides constructive feedback, c) is mostly understanding of the TC role, and d) works with TC and SP, but can sometimes be too directive or absent from the process. Supervisors in this range are categorized as having "subpar quality."

Toward the middle of the variable map are individuals whose score lies in the range of 29 to 30, which puts them below the highest scenario but above the lowest scenarios. Individuals in this range describe their supervisor as about the same as

someone who a) is mostly familiar with policies/procedures, b) provides helpful feedback, c) is supportive of the TC role, and d) generally works with the TC and SP. This range is an indication of a supervisor with an "average quality."

Existing above the "average" supervisor are those individuals who score in the range of 31-35, putting them slightly below the highest scenarios, but above average and far above low quality supervisors. Individuals in this range describe their supervisor as about the same as someone who a) is very knowledgeable of policies/procedures, b) provides lots of constructive feedback, c) is very supportive of the TC role, and d) frequently works with the TC and SP. Supervisors who fall in this range are described as having "great quality."

Finally, those individuals who score in the range of 36 to 37 best align with the highest scenario (S1), and above all other scenarios described below their location. Individuals in this range describe their supervisor as about the same as someone who a) is an active resource and extremely knowledgeable about policies/procedures, b) provides extensive constructive feedback, c) is extremely supportive of the TC role, and d) partners closely with the TC and SP. Supervisors who fall in this range would be described as having "excellent quality."

For those individuals whose score exceedd 38 and up to 45, they indicate their supervisor's quality was better or much better than the highest quality scenario, and are not captured by any of the descriptions across all nine scenarios. They are labeled as "absolute highest quality." All of these score ranges and their interpretations are presented below in Table 4.14.

Table 4.14

USQ Scale Score Interpretations

Score/Range	Perception of Supervisor Quality	General Information	General Traits
38+	Absolute highest quality	TC alumni rates their US's quality as higher than any of the scenarios	*not captured by this set of scenarios—above*
36-37	Excellent quality	TC alumni rates their US's quality as about the same as S1 and S2 and much more than other scenarios	Active resource, extremely knowledgeable about polices/procedures Provides extensive constructive feedback Extremely supportive of TC role Partners closely with TC and SP
31-35	Great quality	TC alumni rates their US's quality as a little lower than S1 and S2, about the same as S3 and S4, and above other scenarios	Very knowledgeable of policies/procedures Provides lots of constructive feedback Very supportive of the TC role Frequently works with TC and SP
29-30	Average quality	TC alumni rates their US's quality as lower than S1 and S2, a little lower than S3-S4, and about the same as S5, and above other scenarios.	Mostly familiar with policies/procedures Provides helpful feedback Supportive of TC role Generally works with TC and SP

24-28	Subpar quality	TC alumni rates their US's quality as much lower than S1 and S2, lower than S3 and S4, a little lower than S5, about the same as S6, and little above all other scenarios.	Mostly aware of policies/procedures Sometimes provide constructive feedback Mostly understanding of TC role Works with TC and SP, but sometimes too directive or absent from process
19-23	Low quality	TC alumni rates their US's quality as much lower than S1, S2, S3, S4, and S5, lower than S6, and about the same as S7 and S8, and slightly above S9.	Aware of few policies/procedures Provides limited constructive feedback Somewhat understanding of TC role Infrequently works with TC and SP, either too commanding or too absent from process
18 and below	Extremely low quality	TC alumni rates their US's quality as much lower than S1, S2, S3, S4, S5, S6, and S7, slightly below S8, and about the same as S9.	Aware of very few policies/procedures Rarely provides useful feedback Not very understanding of TC role Rarely works with TC and SP, either too overbearing or completely absent from process

Cognitive Interviews

Overview of Interview Participants

For the cognitive interviews, all individuals who completed the full administration survey were asked about their interest in participating in an approximately 30 minute 1:1 interview over Zoom to discuss the survey and its contents. Of the 61 who completed the survey, 33 (54.10%) expressed interest to schedule an interview. These 33 individuals were sent emails to schedule interviews, of which 12 replied (36.36%) and scheduled interviews. These 12 interviews took place over Zoom between the teacher candidate alumni and me over the course of three weeks, from mid-December 2021 to early January 2022.

Demographic Information

For the interviews, the same demographic information asked on the survey was inquired again, with the additional question of gender, of which the results are displayed in Table 4.15. All names listed and examined are pseudonyms. Of the teacher candidate alumni, 10 were female, and two were male. They completed their practicum experiences over the last three academic years (i.e. graduating classes of 2019, 2020, 2021), although the year listed is the year they completed their practicum, and not their graduating class year. Eleven of the 12 individuals were alumni of the undergraduate program, whereas one was an alum of a graduate program. Six of the 12 studied elementary education, five studied secondary education, and one studied moderate special needs. Eight of the 12 reported to still be teaching, while four are not currently teaching.

Table 4.15

Demographic Information for TC Alumni Interviews

Name	Gender	Practicum Year	Program Level	Program	Still Teaching
Peyton	F	2018	Undergraduate	Elementary Education	Yes
Chelsea	F	2019	Undergraduate	Elementary Education	Yes
Kelly	F	2019	Undergraduate	Secondary Education	No
Maddie	F	2019	Undergraduate	Secondary Education	Yes
Zachary	M	2019	Undergraduate	Secondary Education	No
Catherine	F	2020	Graduate	Moderate Special Needs	Yes
Harry	M	2020	Undergraduate	Secondary Education	Yes
Savannah	F	2020	Undergraduate	Secondary Education	No
Kylie	F	2021	Undergraduate	Elementary Education	Yes
Mary	F	2021	Undergraduate	Elementary Education	No
Melodie	F	2021	Undergraduate	Elementary Education	Yes
Sophie	F	2021	Undergraduate	Elementary Education	Yes

Interview Findings & Discussion

There were numerous findings found throughout the interview process, which provided interesting takeaways both for this USQ scale, as well as overall insights for those developing scales using the RGS framework. I first layout findings pertaining to the USQ scale, and after for all RGS scales.

USQ Scale Findings

As per the takeaways specific to the USQ scale, I group these into three specific themes: changes made to augment the quality of "action" verbs, the flexibility provided by the middle category, and word/phrases to be changed.

The first theme emerged throughout the interviews. Participants discussed the specific verbs used in each sentence and commented on how the verb utilized was qualified with the use of an adverb or adjective. For example, the first sentence in S4 states "Chris is informed of most of the policies and procedures of the practicum

experience" whereas, in S5, it reads "Taylor is *mostly* familiar with the policies and procedures of the practicum experience" (italics added). Individuals would point to these modifications as the primary distinguisher of the facet quality being presented and consistently stated this helped inform their ultimate response option selection. As an example, Harry stated, "I want to say this is my supervisor. Mostly familiar, can be a useful resource, generally supports me in my role, generally works with my supervisor and me. Like, generally, that is what happened. So I'd say, about the same."

Additionally, participants indicated how these modifications were represented in the most extreme examples, S1 and S9, and frequently commented that these stood out. In reaction to reading S1, Mary remarked that, "Facilitate and co-construct…first of all, co-construct, definitely sounds like a partner. Like, you're in this together…and facilitate, that also stood out to me. Facilitate, makes me think about someone who will be a mediator…they will listen to what you have to say, and show you that you're listening."

Likewise, on the negative spectrum of scenarios, Catherine had a clear reaction while reading S9, remarking, "Oh God…words that stand out to me. Rarely. That is what I think elicited that reaction. And no useful feedback on my lesson plans and reflections…even the worst supervisor…" In fact, nearly all participants audibly laughed while reading S9, and very quickly answered the item (most frequently with the response option, "My supervisor is much better.") Particularly for low quality scenario items, interview participants answered the scenarios much faster than other items, and would cite the severity of the adverb/adjective as the motivator for their fast response selection.

The second theme that consistently emerged was the discussion over the middle response option, "My supervisor and {name} are about the same." Due to the nature of

the number of response options (i.e. five) and the number of scenario items (i.e. nine), individuals will at some point utilize the same response option for multiple scenarios. By far, the most utilized and repeated response option was the middle. Of the 108 responses across the twelve individuals and nine scenarios, 32% of the responses were the middle response option. This was followed in usage by response options 5 (24%), 4 (21%), 2 (13%), and 1 (9%). Many interview participants expressed that the specific phrasing "about the same" provided flexibility in choosing that as their response selection. Maddie touched on this, reflecting, "Maybe my supervisor wasn't working/partnering very closely with my supervising practitioner. But, all of the other things were the same, and I didn't think this scenario any worse or better than mine. It leaves enough wiggle room that you can apply it to your experience."

These findings strongly relate to one of the most discussed issues in scale development research, which is the optimal number of response categories. This is a hotly debated topic, with an incredible amount of disagreement existing across the field. There is a plethora of published literature on the topic, with researchers advocating for 5-point scales (Komorita & Graham, 1965), 7 or even 10-point scales (Alwin & Krosnick, 1991), and even 11-point scales (Alwin, 1997). From a reliability perspective, 7-point scales have been found to maximize reliability (Nunnally, 1967), although other research has suggested that the number of response categories is independent regarding reliability (Matell & Jacoby, 1971). From a validity perspective, one study found 9-point scales to have the highest criterion validity, although those with 5 up through 11-point scales had similar criterion validity (Preston & Coleman, 2000). Thus, the field of scale development highly disagrees with itself in this manner.

The middle response category sometimes referred to as the "neutral point," has been researched and discussed for decades (Guy & Norvell, 1977), and remains controversial. These interviews demonstrated that even when an individual's preferences were in fact slightly tilted either toward their supervisor or the scenario supervisor, the broadness afforded by the phrase "about the same" was listed as the primary reason for selecting that response. One possible approach in future interactions of this scale would be to increase the number of response categories to 7, with the middle category split into "My supervisor is slightly better," "My supervisor and Supervisor X are the same," and "Supervisor X is slightly better."

Finally, participants remarked about specific words/phrases that were possibly confusing, or could be changed. Across all 12 interviews, the use of the word "discouraged" in S9 was acknowledged as contributing to selecting their particular response option. The interview process illuminated that all other items used positive verbiage (e.g. they act, they provide) with adverb/modifications to the amount that verb was enacted (e.g. frequently, mostly). However, the word "discouraged" was used in S9, a departure from the overall pattern of scenario construction. Participants frequently remarked the use of the word "discouraged" stood out as very negative, and this contributed to a "different" feeling than other scenarios. Given this response, I would revise this wordage in future iterations of the scale, perhaps to "rarely encourages me." Interestingly enough, while everyone agreed they would support the change of the wording to this, everyone also stated that this would *not* have changed their ultimate response selection.

Table 4.16 displays the score chart for the twelve participants and it shows their responses for each item, their total score, and the best matching scenario given their total score. For 11 of the 12 interview participants, when they answered the item predicted by their total score to be their best fitting item, they selected the response option "My supervisor and {name} are about the same." This is a strong indication of the scale's overall stability and functionality, as respondents almost universally provided the expected response given their scale total score.

Table 4.16

Cognitive Interviews, Score Chart

Name	S1	S2	S3	S4	S5	S6	S7	S8	S9	Total Score	Best Scenario Match	Rating
Sophie	3	4	4	5	5	5	5	5	5	41	S1 – Alex	Highest
Melodie	3	2	3	4	4	5	5	5	5	36	S2 – Jordan	Excellent
Kylie	3	3	3	3	4	5	4	5	5	35	S3 – Riley	Great
Maddie	3	3	3	3	4	4	5	5	5	35	S3 – Riley	Great
Chelsea	2	2	3	3	4	4	4	5	5	32	S4 – Chris	Great
Kelly	2	3	2	3	4	4	4	5	5	32	S4 – Chris	Great
Mary	2	3	3	3	3	4	4	5	5	32	S4 – Chris	Great
Zachary	1	3	3	3	3	4	4	5	5	31	S4 – Chris	Great
Catherine	2	2	3	3	3	3	4	5	4	29	S5 – Taylor	Average
Harry	1	1	2	2	3	3	3	5	4	24	S6 – Quinn	Subpar
Savannah	1	1	2	4	2	3	3	3	4	23	S7 – Kyle	Low
Peyton	1	1	1	1	1	2	2	3	3	15	S9 – Casey	Extremely Low

At the conclusion of each interview, participants were asked to read the final item on the screen (S4), and to generate a word/phrase for the quality they believed was embodied in each sentence. In essence, participants were asked to generate a facet name as an exercise to compare their responses with my original names for the facets. These results are presented in Table 4.17. There were two overall trends that provided interesting results. First, participants overwhelmingly labeled the facets of "constructive

feedback" and "collaboration" with similar wording of "feedback" or "collaborative," respectively. This is a strong indication that these facets are appropriately named. The second trend was the varied responses given for the facets "resourcefulness" and "mentorship." Participants utilized the words/phrases "policies," "procedures," and "knowledge," and not one individual stated "resourcefulness." In previous facet name iterations, the term "knowledge" was used but was changed to "resourcefulness," as it was noted that supervisors could be knowledgeable, but the true measure of quality is what supervisors *do* with that knowledge (i.e. dispense resources). One possible explanation for this is that the first half of the second sentence utilizes the word resource, while its latter half sometimes explicitly references the action of being a resource. Nevertheless, no respondents labeled the second sentence as "resourcefulness," opting for the term "feedback." However, future iterations of the scale will maintain the name of "resourcefulness." Similarly, although no respondents labeled the third facet as "mentorship" (opting most frequently for the word "support"), this facet was the clearest one that emerged throughout the literature. Thus, given that this scale would be primarily administered and analyzed by teacher educators and practicum directors, I would maintain this as the facet name.

Table 4.17

Cognitive Interviews, Facet Names

ID	Facet 1 Resourcefulness	Facet 2 Constructive Feedback	Facet 3 Mentorship	Facet 4 Collaboration
Sophie	Knowledge	Feedback	Encouragement	Facilitation
Melodie	Policy/ Procedures	Lesson Feedback	Moral Support	Relationship
Kylie	Well-informed/ Trained	Experienced	Encouraging	Collaborative
Maddie	Informed	Feedback	Support	Collaborative
Chelsea	Policies	Feedback	Support	Collaboration
Kelly	Knowledgeable	Feedback/ Help	Understanding	Collaboration
Mary	Policy/ Procedures	Feedback	Support	Collaboration
Zachary	Competence	Feedback	Support	Collaboration
Catherine	Policy	Teacher	Helper	Liaison
Harry	Preparedness	Teacher's Aide	Advocate/ Support	Interaction
Savannah	Expert	Teacher Educator	Coach	Collaborator
Peyton	Knowledge	Feedback	Support	Partnership/ Collaboration

RGS-related Findings

In addition to the specific findings from the USQ scale, participants provided insights about RGS scales, particularly concerning their formatting and presentation. There were three important themes that emerged across the interviews. First, participants universally prefer the method of receiving one scenario item at a time. An oft-repeated word to describe this preference was that it would be "overwhelming" to receive all scenario items at once. During the full administration of the survey, respondents were presented with all scenarios at once, something that will be amended and recommended for future iterations of the scale. Second, participants commented on the randomized

ordering of the scenarios. They state a preference for a pre-determined ordering, rather than complete randomization. In the full administration survey, scenarios were presented in a completely randomized ordering. Additionally, this ordering was different for *each* participant (as a measure to distribute start-up effects across all items). During the interviews, a consistent ordering of scenarios was maintained, in which scenarios of varying degrees of difficulty were distributed across the scale. Given the overall success of the scenario scale and the preference for this pre-determined structure, future iterations would maintain a consistent order across all respondents, and utilize the same ordering as done during the interviews. Further, interview participants support an option to revisit their answers after completing the scale, and the ability to change their answers on one screen. Not all participants agree that they would necessarily revisit their options, but all participants support the notion of providing it as an option. Thus, future iterations of the scale would include an option to revisit all answers on one page, with the ability to change previous responses. Finally, other RGS scales should consider testing the use of a 7-point scale in addition to testing 5-point scales, and analyzing the strengths and weaknesses of their respective psychometric properties.

Summary of Results

During this chapter, I presented the results of the Rasch analyses from both the pilot study and full administration, as well as the results from a series of 12 cognitive interviews. The results of this chapter were presented to answer this dissertation's research questions.

1. *To what extent can a Rasch/Guttman Scenario (RGS) scale development approach be used successfully in the development of a university supervisor scale?*
2. *To what extent does a RGS scale affect the quality of information gained as perceived by survey participants?*

The results from the pilot study provide an initial strong proof of concept, which forms the basis for specific item revisions for the full administration. Word changes and phrases were made in an effort to further differentiate items in the middle of the variable map, due to the proximities of their item difficulties, and the desire to have a more even spread between items. The results from the full administration, specifically the Rasch results demonstrating goodness-of-fit, the CCCs, and an analysis of residuals are strong indications that the answer to research question one is, yes. The RGS scale development approach is successful in constructing a scale to measure the quality of university supervisors. Additionally, the results from the set of cognitive interviews likewise support an affirmative answer to research question two. The use of cognitive interviews provide quality information unique to the interviews and not obtained by the pilot study or full administrations.

For the final chapter, I provide a concluding discussion that builds upon and summarizes the previous chapters that covered the dissertation's purpose, research literature, methodology, and results. Additionally, this final chapter provides insight as to the specifics of how this dissertation contributes to the broader research literature on both teacher education and instrument development. Finally, I discuss the limitations of the study, recommendations for future research, and usage of the scale in practice.

CHAPTER FIVE: DISCUSSION

Throughout the course of the preceding four chapters, I lay out the purpose and need for my dissertation's research, review the fields of literature pertaining to both teacher education, student teaching supervision, as well as the Rasch/Guttman scenario (RGS) scale framework. Additionally, I describe the methodology of my research and ultimately demonstrate in Chapter 4 how I successfully apply the RGS framework to develop an instrument that measures the quality of the university supervisor in student teaching. Furthermore, I display the success of the use of cognitive interviews in providing unique insights when developing an RGS scale. In this concluding chapter, I provide insight on how this research contributes to both the fields of teacher education student teaching supervision, as well as the field of scale development, particularly that of RGS-related scales. Finally, I discuss the limitations of this research and provide recommendations for future research in this area.

Overview of Findings

There are two primary research questions addressed in this dissertation.

1. *To what extent can a Rasch/Guttman Scenario (RGS) scale development approach be used successfully in the development of a university supervisor scale?*
2. *To what extent does a RGS scale affect the quality of information gained as perceived by survey participants?*

The RGS scale development approach was implemented to develop a scale that could measure the perceptions of alumni of teacher education programs regarding the quality of their university supervisor during their student teaching.

The first step in the process was to review the existing literature on RGS instrumentation and in the field of teacher education, and more specifically that of clinical supervision. In addition, after consultation with multiple field placement directors, teacher educators, teacher candidates, and university supervisors, a pilot study was conducted to establish a *proof of concept*. The results of this initial pilot study were encouraging, which showed that the scenarios intended to represent a supervisor with the highest qualities and traits, were in fact the most difficult for people to endorse the highest response option (i.e. "My supervisor is much better than Supervisor X"). Similarly, those scenarios that were intended to represent a supervisor with the lowest qualities and traits, were the easiest for individuals to endorse that same highest response option. Nevertheless, there were certain areas of the scale, particularly the middle of the scale's difficulty level, that were more closely aligned with one another than desired. Additionally, the final item of this middle group (i.e. scenario 6) had a much larger gap between it and the next item, scenario 7. This prompted a round of item revisions, which included introducing more language to better differentiate items to achieve a more uniform spacing of item difficulties.

The results from the final administration, building upon the insights gained from the analysis of the pilot study, provide a welcome improvement to the scale's overall psychometric properties and resolve many of the issues encountered (e.g. clustering of middle-difficulty items, uneven gaps between items). Nevertheless, there are some areas that could be targeted for future improvement. As an example, the middle response option (i.e. "My supervisor and Supervisor X are about the same") was by far the most utilized across all scenarios. The rationale as to why this category may have had such

high utilization was illuminated during the round of 12 cognitive interviews conducted after the final survey administration. Individuals frequently commented about the "flexibility" or "wiggle room" that the "about the same" wording provided them as they debated their ultimate response selection. This suggests that this category is perhaps too broad, and the overall scale might benefit from an expansion to 7 response options from 5. The inclusion of new response options "My supervisor is slightly better" and "Supervisor X is slightly better" as well as a modified neutral category "My supervisor and Supervisor X are the same" may help to rectify this relatively minor psychometric misfit.

Discussion of Findings

There are several contributions that this dissertation's research makes to the fields of teacher education and scale development. First, this research provides a psychometrically stable instrument that measures the perceptions of teacher candidates about their university supervisor's quality. Not only could this scale be utilized by programs for their own program improvement, but also be used as a valid and reliable source of evidence as required by national and state accreditation visits. The lack of available instrumentation for programs to utilize has been an issue, and this research provides a valuable contribution to that field. Additionally, the utilization of the RGS framework provides those in charge of university supervisors with an instrument that provides a much greater level of interpretability in regards to survey results. An individual's total score on the USQ scale places the perception of that supervisor's quality with a vivid, well-rounded description of an individual's practice. For

programmatic improvement, an individual's score also provides directors of field experience with clear, targeted measures to assist in future training of their supervisors.

Additionally, the use of cognitive interviews in conjunction with a RGS scale development process showed to be a beneficial addition. Insights to response patterns (e.g. why the aforementioned middle category was utilized so often) as well as overall preferences and reactions to taking RGS scales (e.g. preference for seeing all items individually, with an option to review at the end) are valuable insights that could not be gained merely from analyzing survey results. The success of the RGS framework in creating a scale with no counterparts in the literature demonstrates the continuing validation of this approach to the field of scale development.

Finally, this dissertation displays the advantages of the RGS framework in scale development over traditional approaches. One of the greatest assets of the RGS framework is the ability to more easily identify areas for item revision. The nature of scenario-based items provides a richer foundation for adjusting language to improve a scale's psychometric properties. The results of the pilot study showed several instances in which items were too closely spaced, a sign of redundancy in the items. The RGS framework allows for more precise identification of problem areas and implements more targeted revisions. This ability to better target problem areas is perhaps one of the strongest arguments for scale developers to utilize the RGS approach.

Limitations

There were many promising results from this dissertation's research, however, it is important to acknowledge some limitations of this study. The primary limitation of this study lies in relation to the representatives of the sample. First, these findings may be

affected by non-response bias, given the large non-response rates of 35.19% and 16.76% on the pilot and full administration, respectively. All respondents were from the same institution, which is a selective private university with a strong teacher education program and is surrounded by some of its state's finest private and public elementary and secondary institutions. Additionally, the model of teacher education at this institution is highly influenced by the policies and practices mandated by the Commonwealth of Massachusetts, which have slight differences from other states across the nation. Finally, all students are supervised by adjunct university supervisors, which while common, is not the only model used by teacher preparation programs. These limitations were discussed prior in Chapters Two and Three; however, it is important to reiterate these limitations as this study's conclusion is discussed.

Additionally, it is important to note the different context that the respondents of the study withstood while in their preparation programs years. In the analysis of alumni perception of teacher education programs, a three-year period from the time of graduation is standard practice. However, the three graduating class years of this study—2019, 2020, 2021—certainly lived during different times. Students of the class of 2019 graduated from their program in a pre-Covid world, while those in the class of 2020 completed their programs during the early days of the pandemic, and the class of 2021 completed their programs also heavily impacted by the pandemic as well.

During March 2020, a rapid transformation of the educational landscape took place in Massachusetts and across the nation and world. Millions of PK-12 students and their teachers were abruptly forced to shift their classrooms to a completely virtual model. These changes had a direct impact on the field of teacher preparation, as student

teachers and their supervisors were forced to complete field experiences in a virtual setting. In Massachusetts, DESE (2020) instituted many new measures for preparation programs, allotting much greater flexibilities for requirements such as program waivers, field experience hours and settings, and virtual observations.

Nevertheless, while important to acknowledge the impact of Covid-19 on education, both at the elementary/secondary levels and in higher education, great attention was paid to the language of these scenarios to be universal in their application. Phrases such as "in person," "visited my classroom," "met with my supervising practitioner and me," are avoided. Instead, words and phrases are used to describe the heart of each facet quality, and not whether these qualities happened in person or over Zoom.

Areas for Future Research

There are multiple avenues that this dissertation's research could provide future research, both in the fields of teacher education and scale development. Specific to the USQ scale, there are certain modifications that could be done to improve the scale even further. As previously mentioned, the expansion of the number of response categories to 7 from 5 might help to alleviate the overuse of the neutral response category. Furthermore, one could study the differences in the scale's stability by administrating the 5-response scale and 7-response scale to the same group and comparing their results. Additionally, slight word changes to individual items might be recommended, such as replacing the negative word "discouraging" with a more positive verb (e.g. encourages), but modified by how often that quality was enacted (e.g. rarely), which is consistent with other scenario sentence construction.

Additionally, it would be interesting to see this scale's implementation among groups of teacher candidates from institutions with different programmatic structures. This includes examining other programs within the Commonwealth of Massachusetts (both adjunct supervisor or full-time clinical professor models), as well as programs outside of Massachusetts, also broken down by adjunct or full-time supervisory models. Language would most likely have to be slightly modified (e.g. Massachusetts uses the term "program supervisor" while most other states use "university supervisor.") It would be interesting to decipher whether the scale's stability holds across institutions of different programmatic models and state oversight.

Lastly, while this dissertation placed its focus on the role of the university supervisor, there is another important member of the student teaching triad: the K-12 classroom teacher (otherwise known as cooperating teacher, supervising practitioner). This individual works with the teacher candidate on a daily basis and is a vital part of the teacher candidate's field experience (Glickman & Bey, 1990). This dissertation focuses on the role of the university supervisor given its lack of attention in the literature, as compared to the supervising practitioner. With the success of this dissertation's research, it would be interesting to see the administration of this scale for supervising practitioners, as well as the construction of a new RGS scale that more specifically targets the unique facets embodied by the supervising practitioner.

Implications and Conclusions

This dissertation demonstrates that the Rasch/Guttman scenario (RGS) framework for scale development can be utilized to measure the perceptions of teacher candidates regarding the quality of their university supervisor. This research builds upon an

emerging field of promising scale development that utilizes the RGS framework. Moreover, my methodology, results, and analysis provide a clear direction for those who seek to use this framework for scale development in their respective fields. Additionally, this dissertation provides a unique contribution to the RGS literature by including the use of cognitive interviews to complement a final scale administration. The illuminating results of the cognitive interviews provide a strong rationale for its use in future RGS scales.

This dissertation concerns itself with the topic of student teaching because it has been shown to be the most influential part of the teacher education program (Cochran-Smith, 1991; Darling-Hammond, 2014; Evertson, 1990). However, despite its purported importance, the research literature is sparse when relating to the role of the university supervisor. Given the importance of this experience, and the impact this individual provides on a teacher candidate, it was important to understand the construct of this role's quality, and how to measure it. This dissertation codifies what truly embodies a supervisor of the highest quality, and builds an instrument that measures where a teacher candidate believes their supervisor lies on that spectrum of quality.

The field of education continues to wrestle with tremendous challenges, especially during the times of Covid-19. The teachers of America are some of our bravest front-line workers, whose impact on our communities stretches far outside the walls of their classroom. As the next generation of teachers enters the workforce, they deserve teacher education programs and clinical experiences that best prepares them for their very important role as educators. I hope that this research will help contribute to improving that experience.

REFERENCES

Allen, J. M., & Wright, S. E. (2014). Integrating theory and practice in the pre-service teacher education practicum. *Teachers and Teaching*, *20*(2), 136-151. https://doi.org/10.1080/13540602.2013.848568.

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)*. Measured Progress and SRI International.

Alwin, D. F. (1997). Feeling thermometers versus 7-point scales. Which are better? *Sociological Methods & Research*, *25*(3), 318-340.

Alwin, D. F., & Krosnick, J. A. (1991). The reliability of survey attitude measurement: the influence of question and respondent attributes. *Sociological Methods & Research*, *20*(1), 139-181.

Anderson, J. D. (1988). The education of blacks in the South, 1860–1935. The University of North Carolina Press.

Anderson, D. J., Major, R. L., & Mitchell, R. R. (1992). *Teacher supervision that works: A guide for university supervisors*. Praeger Publishers.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561-573.

Antipkina, I., & Ludlow, L. H. (2020). Parental involvement as a holistic concept using Rasch/ Guttman scenario scales. *Journal of Psychoeducational Assessment*, 00(0), 1-20.

Asplin, K. N., & Marks, M. J. (2013). Increasing the influence of university supervisors during student teaching. *Teaching and Teacher Education, 37*(1), 237-342.

Association for Student Teaching. (1964). *The college supervisor: Conflict and challenge.* WM. C. Brown Co., Inc.

Baez-Cruz, M. (2019, April 25). Sociocultural integration. In L.H. Ludlow (Chair), *Facet design and Rasch measurement: An innovative approach to instrument development* [Symposium] New England Educational Research Organization, Portsmouth, NH, United States.

Beatty, P. C. (2003). *Answerable questions: Advances in the methods for identifying and resolving questionnaire problems in survey research* (Publication No. 3106016). [Doctoral dissertation, University of Michigan]. ProQuest Dissertations Publishing.

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly*, *71*(2), 287-311.

Beck, C., & Kosnik, C. (2002). Components of a good practicum placement: Student teacher perceptions. *Teacher Education Quarterly*, *29*(2), 81–98.

Bishop, G. (1987). Experiments with the middle response alternative in survey questions. *Public Opinion Quarterly*, *51*, 220–232.

Bobbitt, J. F. (1913). Some general principles of management applied to the problems of city-school systems. In J.F. Bobbitt (Ed.). *Twelfth yearbook of the national society for the study of education, Part I: The supervision of city schools* (pp. 7-96). The University of Chicago Press.

Borg, I., & Shye, S. (1995). *Facet theory*. Thousand Oaks, CA: Sage.

Briggs, F. M. (1963). The unique and complex role of the college supervisor of student teachers. *The High School Journal*, *46*(8), 291-295.

Byrd, D., & Fogelman, J. (2012). The role of supervision in teacher development. In A. Cuenca (Ed.), *Supervising student teachers: Issues, perspectives, and future directions* (pp. 191-210). Sense Publishers.

Callahan, R. E. (1962). *Education and the cult of efficiency*. University of Chicago.

Chang, W-C. (2017). *Measuring the complexity of teachers' enactment of practice for equity: A Rasch model and facet theory-based approach* (Publication No. 10269093) [Doctoral dissertation, Boston College]. ProQuest Dissertations Publishing.

Chang, W-C. (2019, April 25). Teachers' equity practices. In L.H. Ludlow (Chair), *Facet design and Rasch measurement: An innovative approach to instrument development* [Symposium] New England Educational Research Organization, Portsmouth, NH, United States.

Chang W-C, Ludlow L. H., Grudnoff, A., Ell, F., Haigh, M., Hill, M., & Cochran-Smith, M. (2019). Measuring the complexity of teaching practice for equity: Development of a scenario-format scale. *Teaching and Teacher Education*, *82*, 69-85.

Christophersen, K. A., Elstad, E., Solhaug, T., & Turmo, A. (2016). Antecedents of student teachers' affective commitment to the teaching profession and turnover intention. *European Journal of Teacher Education*, *39*(3), 270-286.

Clifford, G.J., & Guthrie, J.W. (1990). *Ed school: a brief for professional education*. The University of Chicago Press.

Cochran-Smith, M. (1991). Reinventing student teaching. *Journal of Teacher Education*, *42*, 104-118.

Cochran-Smith, M. (1999). Learning to teach for social justice. In G. Griffin (Ed.), The education of teachers: *Ninety-eighth yearbook of the national society for the study of education* (pp. 114-144). University of Chicago Press.

Cochran-Smith, M. (2003). Learning and unlearning: the education of teacher educators. *Teaching and Teacher Education*, *21*(3), 5-28.

Cochran-Smith, M. (2005). The new teacher education: For better or for worse? *Educational Researcher*, *34*(7), 3-17.

Cochran-Smith, M. (2010). Toward a theory of teaching education for social justice. In M. Fullan, A. Hargreaves, D. Hopkins & A. Lieberman (Eds.), *The international handbook of educational change* (2nd ed.). Springer Publishing.

Cochran-Smith, M., & Dudley-Manning, C. (2001). The flunk heard round the world. *Teaching Education*, *12*(1), 49-63.

Cochran-Smith, M., Keefe, E. S., Carney, M. C., Olivo, M., & Jewett Smith, R. (2020). Teacher preparation at new graduate schools of education: Studying a controversial innovation. *Teacher Education Quarterly*, *47*(2), 8-37.

Cochran-Smith, M., Piazza, P., & Power, C. (2013). The politics of accountability: Assessing teacher education in the United States. *The Educational Forum, 77*(1), 6-27.

Cochran-Smith, M., & Power, C. (2010). *New Directions for Teacher Preparation*. ASCD. https://www.ascd.org/el/articles/new-directions-for-teacher-preparation

Cochran-Smith, M. & Villegas, A. M. (2014). Framing teacher preparation research: An overview of the field, part 1. *Journal of Teacher Education*, *65*(4), 7-20.

Conant, J. B. (1963). *The education of American teachers*. McGraw-Hill.

Cook, R. M., McKibben, W. B., & Wind, S. A. (2018). Supervision perception of power in clinical supervision: the power dynamics in supervision scale. *Training and Education in Professional Psychology*, *12*(3), 188-195.

Council for the Accreditation of Educator Preparation. (2016). *Council for the Accreditation of Educator Preparation report to the public, the states, the policymakers, and the education profession*. http://www.caepnet.org/standards/introduction

Cowan, J., Goldhaber, D., Jin, Z., & Theobald, R. (2020). *Teacher licensure tests: Barrier or predictive tool?* CALDER. https://caldercenter.org/publications/teacher-licensure-tests-barrier-or-predictive-tool.

Cremin, L. A. (1957). *The republic and the school: Horace Mann on the education of free men*. Teachers College Press.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart, and Winston, Inc.

Crotty, T., & Allyn, D. (2001). Evaluating student reflections (ED459174). ERIC. http://files.eric.ed.gov/fulltext/ED459174.pdf

Cuenca, A. (2010). Care, thoughtfulness, and tact: a conceptual framework for university supervisors. *Teaching Education*, *21*(3). 263-278.

Cuencaa. A., Schmeichelb, M., Butlerd, B. M., Dinkelmanb, T., & Nichols, J. R. (2011). Creating a "third space" in student teaching: Implications for the university supervisor's status as outsider. *Teaching and Teacher Education*, 27, 1068-1077.

Darling-Hammond, L. (2014). Strengthening clinical preparation: the holy grail of teacher education, *Peabody Journal of Education*, *89*, 547-561.

Darling-Hammond, L., Burns, D., Campbell, C., Goodwin, A. L., Hammerness, K., Low, E. L., & Zeichner, K. (2017). *Empowered educators: How high-performing systems shape teaching quality around the world*. John Wiley & Sons.

Dewey, J. (1929). *The sources of a science of education*. Liveright.

Donovan, M. K., & Cannon, S. O. (2018). The university supervisor, edTPA, and the new making of the teacher. *Education Policy Analysis Archives*, *26*(28). http://dx.doi.org/10.14507/epaa.26.2849

Ediger, M. (2009). Supervising the student teacher in public school. *Education*, *130*(2), 251–254.

Ellis, N., & Osborne, S. (2015). Mentoring-collaborative approach. *Independent Education*, *45*(2), 14.

Enterline, S., Cochran-Smith, M., Ludlow, L. H., & Mitescu, E. (2008). Learning to teach for social justice: measuring change in the beliefs of teacher candidates. *The New Educator*, 267-290.

Enz. B. J., Freeman, D. J., & Wallin, M. B. (1996). Roles and responsibilities of the student teacher supervisor: matches and mismatches in perception. In D. J. McIntyre & D. M. Byrd (Eds.), *Preparing tomorrow's leaders: the field experience: teacher education yearbook IV* (pp. 131-150). Corwin Press.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. MIT Press.

Evertson, C. M. (1990). Bridging knowledge and action through clinical experiences. In D. D. Dill (Ed.), *What teachers need to know* (pp. 94-109). Jossey-Bass.

Fowler, R. C. (2001). What did the Massachusetts teacher tests say about American education? *Phi Delta Kappan*, *82*(10), 773-780.

Framingham State University. (2021). *History and past presidents*. https://www.framingham.edu/about-fsu/presidents-office/presidential-inauguration/history-and-past-presidents/index

Fraser, J. (2007). *Preparing America's teachers: a history*. Teachers College Press.

Freeman, D. (1990). Intervening in practice teaching. In J.C. Richards & D. Nunan (Eds.), *Second language teacher education* (pp. 103-117). Cambridge University Press.

Friborg, O., Martinussen, M., & Rosenvinge, J. H. (2006). Likert-based vs. semantic differential-based scorings of positive psychological constructs: A psychometric comparison of two versions of a scale measuring resilience. *Personality and Individual Differences*, *40*, 773–844.

Ganser, T. (1996). The coopering teacher role. *Teacher Education*, *31*(4), 283-291.

Glanz, J., & Hazi, H. M. (2019). Shedding light on the phenomenon of supervision traveling incognito: A field's struggles for visibility. *Journal of Educational Supervision*, *2*(1).

Glickman, C. D., & Bey, T. M. (1990). Supervision. In Handbook of research on teacher education: A project of the association of teacher educators (pp. 549-568). MacMillan Publishing Company.

González-Toro, C. M., Cherubini, J.M., Doig, S. R., & Fernández-Vivó, M. (2020). Supervisor feedback: Perceptions from physical education teacher candidates. *The Physical Educator*, *77*, 553-574.

Guttman, L. (1954). An outline of some new methodology for social research. *Public Opinion Quarterly, 18*, 395-404.

Guttman, L. (1959). Introduction to facet design and analysis. In *Proceedings of the Fifteenth International Congress of Psychology* (pp. 130-132). North Holland.

Guttman, R., & Greenbaum, C. W. (1998). Facet theory: Its development and current status. *European Psychologist, 3*(1), 13-36.

Guy, R. F., & Norvell, M. (1977). The neutral point on a Likert scale. *The Journal of Psychology*, *95*, 199-204.

Guyton, E., & McIntyre, J. (1990). Student teaching and school experiences. In W. R. Houston (Ed.), *The handbook of research on teacher education* (pp. 514-534). MacMillan.

Haines, A. (1960). *Building the Student Teaching Process in Elementary Education*. Rand McNally.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analyses* (7th ed.). Pearson.

Hambleton, R. K., & Cook, L. (1977). Latent trait models and their use in analysis of educational test data. *Journal of Educational Measurement*, *14*(2), 75-96.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications.

Henry, G. T., Kershaw, D. C., Zulli, R. A., & Smith, A. A. (2012). Incorporating teacher effectiveness into teacher preparation program evaluation. *Journal of Teacher Education*, *63*(5), 335-355.

Higher Education Research Institute. (2020). 2020 CIRP freshman survey. https://ucla.app.box.com/v/TFS-instrument.

Hobson, A. J., Ashby, P., Malderez, A., & Tomlinson, P. D. (2009). Mentoring beginning teachers: What we know and what we don't. *Teaching and Teacher Education*, *25*(1), 207–216.

Hogue, K. (2019, April 25). Teachers' data use. In L.H. Ludlow (Chair), *Facet design and Rasch measurement: An innovative approach to instrument development* [Symposium] New England Educational Research Organization, Portsmouth, NH, United States.

Holbrook, K. R. (2019, April 25). Student teacher supervision quality. In L.H. Ludlow (Chair), *Facet design and Rasch measurement: An innovative approach to instrument development* [Symposium] New England Educational Research Organization, Portsmouth, NH, United States.

Hosic, J. F. (1920). The democratization of supervision. *School and Society*, *11*, 331-336.

Hudson, P. (2016). Forming the mentor–mentee relationship. *Mentoring & Tutoring: Partnership in Learning*, *24*(1), 30–43.

Humphry, S. (2002). Residuals and rating scales. *Rasch Measurement Transactions*, *16*(1), 866.

Hunter, M. (1973). Appraising teaching performance: One approach. *The National Elementary Principal*, 55(2), 60-62.

Izadinia, M. (2015). Talking the talk and walking the walk: Pre-service teachers' evaluation of their mentors. *Mentoring & Tutoring: Partnership in Learning*, *23*(4), 341-353.

Izadinia, M. (2016). Preservice teachers' professional identity development and the role of mentor teachers. *International Journal of Mentoring and Coaching in Education*, *5*(2), 127-143.

Johnson, I. L., & Napper-Owen, G. (2011). The importance of role perceptions in the student teaching triad. *Physical Educator*, *43*, 46-56.

Kalton, G., Roberts, J., & Hold, D. (1980). The effects of offering a middle response option with opinion questions. *Journal of the Royal Statistical Society Series D (the Statistician)*, *29*, 65–78.

King, S. E. (2008). Inspiring critical reflection in preservice teachers. *The Physical Educator*, *65*(1), 21–29.

Koerner, M., & Rust, F. O. (2002). Exploring roles in student teaching placements. *Teacher Education Quarterly*, *29*(2), 35-53.

Komorita, S. S., & Graham, W. K. (1965). Number of scale points and the reliability of scales. *Educational and Psychological Measurement*, *25*(4), 987-995.

Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In J. D. Wright & P. V. Marsden (Eds.), *Handbook of survey research* (2nd ed., pp. 263-314). Emerald Group.

LeGeros, L. (2013). *The association between elementary teacher licensure test scores and student growth in mathematics: An analysis of Massachusetts MTEL and MCAS tests*. (Publication No. 3608303) [Doctoral dissertation, University of Massachusetts Boston]. ProQuest Dissertations Publishing.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean. *Rasch Measurement Transactions*, *16*, 878.

Linacre, J. M. (2022a). Dimensionality: contrasts & variances. https://www.winsteps.com/winman/principalcomponents.htm

Linacre, J. M. (2022b). Reliability and separation of measures. https://www.winsteps.com/winman/reliability.htm

Ludlow, L. H. (1983). *The analysis of Rasch model residuals*. [Unpublished doctoral dissertation]. University of Chicago.

Ludlow, L. H. (2001). Teacher test accountability: from Alabama to Massachusetts. *Educational Policy Analysis Archives*, *9*(6), 1-22.

Ludlow, L. H. (2017). Rasch: analysis of fit. [Classroom presentation]. Boston College, Chestnut Hill, MA, United States.

Ludlow L. H., Anghel, E., Szendey, O., O'Keefe, T., Howell, B., Matz-Costa, C., & Braun, H. (2020). The Boston College Living a Life of Meaning and Purpose (BC-LAMP) Portfolio: An application of Rasch/Guttman Scenario methodology. *Journal of Applied Measurement*, *21*(2), 134-53.

Ludlow, L. H., Baez-Cruz, M., Chang, W-C., & Reynolds, K. A. (2020). Rasch/Guttman scenario (RGS) scales: A methodological *framework*. *Journal of Applied Measurement*, *21*(4), 361-378.

Ludlow, L. H., Enterline, S., & Cochran-Smith, M. (2008). Learning to teach for social justice: An application of Rasch measurement principles. *Measurement and Evaluation in Counseling and Development*, 194-214.

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.

Ludlow, L. H., Matz-Costa, C., & Klein, K. (2019). Enhancement and validation of the Productive Engagement Portfolio-Scenario (PEP-S8) Scales. *Measurement and Evaluation in Counseling and Development*, *52*(1), 15-37. https://doi.org/10.1080/07481756.2018.1497430

Ludlow, L. H, Pedulla, J., Enterline, S., Cochran-Smith, M., Loftus, F., Salomon-Fernandez, Y., & Mitescu, E. (2008). From students to teachers: using surveys to build a culture of evidence and inquiry. *European Journal of Teacher Education*, *31*(4), 319–337.

Ludlow, L. H., Reynolds, K. A., Baez-Cruz, M., & Chang, W-C. (in press). *Enhancing the interpretation of scores through Rasch/Guttman scenario scales*.

Killian, J. E., & Post, D. M. (1998). The scientific dimensions of supervision. In G. R. Firth & E. F. Pajak (Eds.), *Handbook of research on school supervision* (pp. 1032-1054). Simon & Schuster Macmillan.

Markowitz, R. J. (1993). My daughter, the teacher: Jewish teachers in New York city schools. Rutgers University Press.

Marks, M. J. (2002). *From coursework to classroom: A qualitative study on the influences of preservice teacher socialization*. (Publication No. 3062054) [Doctoral dissertation, University of Cincinnati]. ProQuest Dissertations Publishing.

Martínez-Agudo, J. D. (2016). What type of feedback do student teachers expect from their school mentors during practicum experience? The case of Spanish EFL student teachers. *Australian Journal of Teacher Education*, *41*(5), 36-51.

Massachusetts Department of Elementary and Secondary Education. (2015). *Guidelines for the professional standards for teachers*. http://www.doe.mass.edu/edprep/advisories/TeachersGuidelines.pdf

Massachusetts Department of Elementary and Secondary Education. (2016). *2015-2016 Statewide summary of results, analysis and trends*. http://www.doe.mass.edu/edprep/surveys/surveysummary.docx

Massachusetts Department of Elementary and Secondary Education. (2017a). *Educator preparation surveys: technical report*. http://www.doe.mass.edu/edprep/surveys/2017TechnicalReport.pdf

Massachusetts Department of Elementary and Secondary Education. (2017b). *Program approval criteria list*. http://www.doe.mass.edu/edprep/evaltool/2017CriteriaList.pdf

Massachusetts Department of Elementary and Secondary Education. (2017c). *Educator preparation: 2017 field feedback results*. https://www.doe.mass.edu/edprep/domains/improvement/field-feedback/2017report.docx

Massachusetts Department of Elementary and Secondary Education. (2019). *Guidelines for the candidate assessment of performance: assessment of teacher candidates*. https://www.doe.mass.edu/edprep/cap/guidelines.docx

Massachusetts Department of Elementary and Secondary Education. (2020). *2020-21 ed prep advisory: COVID-19-related impacts on ed prep in 2020-2021*. https://www.doe.mass.edu/edprep/resources/guidelines-advisories/edprep-advisory-2021.pdf

Massachusetts Department of Elementary and Secondary Education. (2022). *Organization search: educator preparation program provider (EPPP)*. https://profiles.doe.mass.edu/search/search.aspx?leftNavId=11238

Massachusetts General Laws c. 69, § 1B; c. 69, §§ 1J and 1K, as amended by St. 2010; c. 12, § 3; c. 71, §§ 38G, 38G ½; c. 76, §19.

Matell, M. S., & Jacoby, J. (1971). Is there an optimal number of alternatives for Likert scale items? Study I: reliability and validity. *Educational and Psychological Measurement*, *31*(3), 657-674.

McIntyre, D. J., & Byrd, D. M. (1998). Supervision in teacher education. In G. R. Firth & E. F. Pajak (Eds.), *Handbook of research on school supervision* (pp. 409-427). Simon & Schuster Macmillan.

McIntyre, D. J., & Killian, J. E. (1987). The influence of supervisory training for cooperating teachers on preservice teachers' development during early field experiences. *Journal of Educational Research*, *80*(5), 277–282.

McNeil, J. D. (1982). A scientific approach to supervision. In T. J. Sergiovanni (Ed.). *Supervision of teaching*. (ASCD 1982 yearbook) (pp. 18-34). ASCD.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). Macmillan.

Metcalf, K. K. (1991). The supervision of student teaching: A review of research. *Teacher Education*, *26*, 27–42.

Miller, A. F. (2017). Creating jaw-droppingly effective rookie teachers: unpacking teacher preparation at the Sposato Graduate School of Education (Match Education) (Publication No. 10266957) [Doctoral dissertation, Boston College]. ProQuest Dissertations Publishing.

Morris, J. E. (1980). Evaluating the effectiveness of the university supervisor of student teachers: role of the coordinator of field experiences. *Peabody Journal of Education*, *57*(2), 148-151.

Napper-Owen, G., & McCallister, S. (2005). What elementary physical education student teachers observe and reflect upon to assist their instruction. *The Physical Educator*, *62*(2), 76–84.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Washington, DC: U.S. Government Printing Office.

National Council for Accreditation of Teacher Education. (2008). Professional standards for the accreditation of schools, colleges, and departments of education. http://www.ncate.org/documents/standards/NCATE%20Standards%202008.pdf

National Council for Accreditation of Teacher Education. (2014). State partnership program FAQ.

http://www.ncate.org/States/NCATEStatePartnershipProgram/StatePartnershipProgramFAQ/tabid/219/Default.aspx#stfaqs9

National Council on Teacher Quality. (2021, March). *State of the states 2021: teacher preparation policy*. https://www.nctq.org/dmsView/PrintReadyStateoftheStates2021TeacherPreparationPolicy

Nunnally, J. C. (1967). *Psychometric theory*. McGraw Hill.

Ogren, C. A. (2013). The history and historiography of teacher preparation in the United States: A synthesis, analysis, and potential contributions to higher education history. In M. B. Paulsen (Ed.), *Higher education: Handbook on theory and research* (28th ed., pp. 405-458). Springer.

Paige, R., Stroup, S., & Andrade, J. R. (2002). *Meeting the highly qualified teachers challenge: The Secretary's annual report on teacher quality*. Washington, DC: U.S. Department of Education Office of Postsecondary Education. http://www.title2.org/ADATitleIIReport2002.pdf

Paige, R., Stroup, S., & Andrade, J. R. (2003). *Meeting the highly qualified teachers challenge: The Secretary's annual report on teacher quality*. Washington, DC: U.S. Department of Education Office of Postsecondary Education. http://www.title2.org/TitleIIReport03.pdf

Pajak, E. (1993). Change and continuity in supervision and leadership. In G. Cawelti (Ed.), *Challenges and achievements of American education* (pp. 158-186). Association for Supervision and Curriculum Development.

Pavan, B. N. (1985). *Hunter's clinical supervision and instruction models: research in schools utilizing comparative measures*. Paper presented at Council of Professors of Instructional Supervision, Washington, D.C.

Payne, K. A. (2018). Democratic teachers mentoring novice teachers: Enacting democratic practices and pedagogy in teacher education. *Action in Teacher Education*, *40*(2), 133-150.

Potter, J. P., Hollas, T., & Coyne, J. (2016). Strengthening collaboration for successful field experiences. *Cogent Education*, 3, 1-30. https://doi.org/10.1080/2331186X.2016.1226561

Power, C. L. (2020). *Up to the task? A policy analysis of the Massachusetts edTPA pilot and field test*. (Publication No. 28261821) [Doctoral dissertation, Boston College]. ProQuest Dissertations Publishing.

Pressley, D. S. (1998a, June 26). Dumb struck: Finneran slams `idiots' who failed teacher tests. *Boston Herald*, 1.

Pressley, D. S. (1998b, June 28). Teacher crisis seen as sad: years of low pay, standards said to add up to low scores. *Boston Herald*, 6.

Preston, C. C., & Coleman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, *104*(2000), 1-15.

Randall, J., & Engelhard, J. G. (2010). Using Guttman's mapping sentences and many facet Rasch measurement theory to develop an instrument that examines the grading philosophies of teachers. *Journal of Applied Measurement*, 11, 122-141.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*.

(Copenhagen, Danish Institute for Educational Research), expanded edition 1980) with foreword and afterword by B. D. Wright. The University of Chicago Press.

Reagan, E. M. (2011). Examining the relationships among undergraduate teacher candidates' experiences, perceptions and beliefs about teaching for social justice. (Publication No. 3454157) (Doctoral dissertation, Boston College]. ProQuest Dissertations Publishing.

Reynolds, K. A. (2019, April 25). Faculty availability outside class. In L.H. Ludlow (Chair), *Facet design and Rasch measurement: An innovative approach to instrument development* [Symposium] New England Educational Research Organization, Portsmouth, NH, United States.

Renolyds, K. A. (2020). *Measuring students' perceptions of faculty availability outside of class using Rasch/Guttman scenario scales*. [Unpublished doctoral dissertation]. Boston College.

Roskam, E. E., & Broers, N. (1996). Constructing questionnaires: An application of facet design and item response theory to the study of lonesomeness. In G. Engelhard & M. Wilson (Eds.), *Objective measurement: Theory into practice*, (3rd ed., pp. 349-385). Able.

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.

Saarikoski, M. (2014). The supervision scale measurement of the clinical learning environment components in a nursing context. In C.E. Watkings & D.L. Milne

(Eds.), *The Wiley international handbook of clinical supervision* (1st ed., pp. 416-430). John Wiley & Sons.

Scheeler, M. C., Ruhl, K. L., & McAfee, J. K. (2004). Providing performance feedback to teachers: A review. *Teacher Education and Special Education*, *27*(4), 396–407.

Schleicher, A. (2011). *Building a high-quality teaching profession: Lessons from around the world*. OECD Publishing.

Sergiovanni, T.J & Starratt, R.J. (1988). *Supervision: Human perspectives* (4th ed.). McGraw-Hill Book Company.

Sergiovanni, T. J. & Starratt, R. J. (1993). *Supervision: Human perspectives* (5th ed.). McGraw-Hill Book Company.

Sergiovanni, T. J. & Starratt, R. J. (2007). *Supervision: Human perspectives* (8th ed.). McGraw-Hill Book Company.

Sharp, C. (1990). Supervision of student teachers: The role of the college supervisor. *Education*, *111*(1), 53-56.

Slick, S. K. (1997). Assessing versus assisting: the supervisor's roles in the complex dynamics of the student teaching triad. *Teaching and Teacher Education*, *13*(7), 713-726.

Steadman, S. C., & Brown, S. D. (2011). Defining the job of university supervisor: a department-wide study of university supervisors' practices. *Issues in Teacher Education*, *20*, 51-68.

Stitzlein, S. M., & West, C. K. (2014). New forms of teacher education: connections to charter schools and their approaches. *Democracy and Education*, *22*.

Taylor, F. W. (1911). *The principles of scientific management*. Harper and Brothers.

Tanner, D., & Tanner, L. (1987). *Supervision in education: Problems and practices*. Macmillan Publishing Company.

Truebner, M. (2019). The dynamics of "neither agree nor disagree" answers in attitudinal questions. *Journal of Survey Statistics and Methodology*, *0*, 1-22.

U.S. Department of Education. (2009). *Race to the Top executive summary*. http://www2.ed.gov/programs/racetothetop/executive-summary.pdf

U.S. Department of Education. (2015). *Every Student Succeeds Act (ESSA): A new education law*. http://www. ed.gov/essa

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Sage.

Wilson, S. M., Floden, R. E., & Ferrini-Mundy, J. (2002). Teacher preparation research: an insider's view from the outside. *Journal of Teacher Education*, *53*(3), 190-204.

Winstanley, J., & White, E. (2014). The Manchester clinical supervision scale© MCSS-26©. In C.E. Watkings & D.L. Milne (Eds.), *The Wiley international handbook of clinical supervision* (1st ed., pp. 386-401). John Wiley & Sons.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement. 8*(2), 97-116.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*, 370-371.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. MESA Press.

Zeichner, K. M. (2005). Becoming a teacher educator: A personal perspective. *Teaching and Teacher Education*, *21*(2), 117-124.

Zeichner, K. M. (2016). Independent teacher education programs: apocryphal claims, illusory evidence. *National Educational Policy Center*.

Zeichner, K. M., & Gore, J. M. (1990). Teacher socialization. In W. R. Houston (Ed.), *The handbook of research on teacher education* (329-348). MacMillan.

Zimpher, N. L., deVoss, G. G., & Nott, D. L. (1980). A closer look at university student teacher supervision. *Journal of Teacher Education*, *31*(4), 11-15.

APPENDIX A: PILOT SCENARIO ITEMS

Scenario 1: Alex is extremely knowledgeable about all of the policies and procedures of the practicum experience. They provide me with vast amounts of constructive feedback on all of my lesson plans and reflections. Alex continuously encourages me in my unique role as a student teacher in my school. They facilitate dialogue between with my supervising practitioner and me to co-construct my practicum experience.

Scenario 2: Jordan is well-informed about the policies and procedures of the practicum experience. They provide me with extensive constructive feedback on my lesson plans and reflections. Jordan encourages me in my unique role as a student teacher. They partner closely with my supervising practitioner and me throughout my practicum experience.

Scenario 3: Riley is informed about the policies and procedures of the practicum experience. They provide me with substantial constructive feedback on my lesson plans and reflections. Riley is considerate of my unique role as a student teacher at my school. They collaborate with my supervising practitioner and me during my practicum experience.

Scenario 4: Chris is familiar with most of the policies and procedures of the practicum experience. They provide me with considerable amounts of valuable feedback on my lesson plans and reflections. Chris is very supportive of my role as a student teacher at my school. They work alongside my supervising practitioner and me during my practicum experience.

Scenario 5: Taylor is familiar with policies and procedures of the practicum experience. They provide me with helpful feedback on my lesson plans and reflections. Taylor supports me in my role as a student teacher at my school. They work with my supervising practitioner and me during my practicum experience.

Scenario 6: Quinn is aware of most policies and procedures of the practicum experience. They provide me with some helpful feedback on my lesson plans and reflections. Quinn is understanding of my role as a student teacher at my school. They work with my supervising practitioner and me, but sometimes they seem to direct the process.

Scenario 7: Kyle is unaware of some policies and procedures of the practicum experience. They provide me with selected amounts of useful feedback on my lesson plans and reflections. Kyle is somewhat understanding of my role as student teacher at my school. They sometimes work with my supervising practitioner and me, but they also can take too commanding a role in the process.

Scenario 8: Sam is aware of few policies and procedures of the practicum experience. They provide me with limited amounts of useful feedback on my lesson plans and reflections. Sam is mostly understanding of my role as student teacher at my school. They infrequently work with my supervising practitioner and me, often taking too dominant a role in the process.

Scenario 9: Casey is aware of very few of the policies and procedures of the practicum experience. They provide me with limited to no useful feedback on my lesson plans and reflections. Casey is slightly understanding of my role as a student teacher at my school. They rarely work with my supervising practitioner and me, frequently taking an overbearing role in the process.

APPENDIX B: FULL ADMINISTRATION SCENARIO ITEMS

Scenario 1: Alex is extremely knowledgeable about all the policies and procedures of the practicum experience. They continuously act as a vital resource, frequently providing me with vast amounts of constructive feedback on all my lesson plans and reflections. Alex always encourages me in my unique role as a student teacher in my school. They facilitate dialogue between my supervising practitioner and me to co-construct my practicum experience.

Scenario 2: Jordan is well-informed about the policies and procedures of the practicum experience. They frequently act as an important resource, providing me with extensive constructive feedback on my lesson plans and reflections. Jordan encourages me in my unique role as a student teacher. They partner closely with my supervising practitioner and me throughout my practicum experience.

Scenario 3: Riley is well-informed about the policies and procedures of the practicum experience. They often act as a valuable resource, providing me with substantial constructive feedback on my lesson plans and reflections. Riley is very considerate of my unique role as a student teacher at my school. They collaborate with my supervising practitioner and me during my practicum experience.

Scenario 4: Chris is informed of most of the policies and procedures of the practicum experience. They act as an important resource, providing me with considerable amounts of valuable feedback on my lesson plans and reflections. Chris is supportive of my role as a student teacher at my school. They work alongside my supervising practitioner and me during my practicum experience.

Scenario 5: Taylor is mostly familiar with the policies and procedures of the practicum experience. They can be a useful resource, providing me with helpful feedback on my lesson plans and reflections. Taylor supports me in my role as a student teacher at my school. They generally work with my supervising practitioner and me during my practicum experience.

Scenario 6: Quinn is mostly aware of the policies and procedures of the practicum experience. They sometimes act as a resource, providing me with some helpful feedback on my lesson plans and reflections. Quinn is mostly understanding of my role as a student teacher at my school. They work with my supervising practitioner and me, but sometimes they seem to direct the process or provide little input.

Scenario 7: Kyle is aware of some, but not all, policies and procedures of the practicum experience. They sporadically act as a resource, providing me with selected amounts of useful feedback on my lesson plans and reflections. Kyle is normally understanding of my role as a student teacher at my school. They sometimes work with my supervising practitioner and me, but they either take too commanding a role, or an insufficient amount, in the process.

Scenario 8: Sam is aware of few policies and procedures of the practicum experience. They infrequently act as a resource, providing me with limited amounts of useful feedback on my lesson plans and reflections. Sam is somewhat understanding of my role as a student teacher at my school. They infrequently work with my supervising practitioner and me, either taking too dominant a role, or are too removed from the process.

Scenario 9: Casey is aware of very few policies and procedures of the practicum experience. They rarely act as a resource, providing me with limited to no useful feedback on my lesson plans and reflections. Casey is not understanding and can be discouraging in my role as a student teacher at my school. They rarely work with my supervising practitioner and me, either frequently taking an overbearing role, or are completely absent from the process.

APPENDIX C: SCENARIO SCORING CHART, BY FACET AND LEVEL

Levels	FACETS			
	Resourcefulness	**Constructive Feedback**	**Mentorship**	**Collaboration**
1	Professional	Extensive	Advocate	Co-constructs with SP, TC
2	Familiar	Selected	Supportive	Collaborates with SP, TC
3	Unaware	Limited	Discouraging	Dictates to SP, TC *or* Absent

Scenario	FACETS				Total Score
	Resourcefulness	**Constructive Feedback**	**Mentorship**	**Collaboration**	
1	3	3	3	3	12
2	3	3	3	2	11
3	3	3	2	2	10
4	2	3	2	2	9
5	2	2	2	2	8
6	2	2	2	1	7
7	1	2	2	1	6
8	1	1	2	1	5
9	1	1	1	1	4

Scenario	Name	FACETS								Total Score
		Resourcefulness		Constructive Feedback		Mentorship		Collaboration		
1	Alex	3	Professional	3	Extensive	3	Advocate	3	Co-constructs with SP, TC	12
2	Jordan	3	Professional	3	Extensive	3	Advocate	2	Collaborates with SP, TC	11
3	Riley	3	Professional	3	Extensive	2	Supportive	2	Collaborates with SP, TC	10
4	Chris	2	Familiar	3	Extensive	2	Supportive	2	Collaborates with SP, TC	9
5	Taylor	2	Familiar	2	Selected	2	Supportive	2	Collaborates with SP, TC	8
6	Quinn	2	Familiar	2	Selected	2	Supportive	1	Dictates to SP, TC *or* Absent	7
7	Kyle	1	Unaware	2	Selected	2	Supportive	1	Dictates to SP, TC *or* Absent	6
8	Sam	1	Unaware	1	Limited	2	Supportive	1	Dictates to SP, TC *or* Absent	5
9	Casey	1	Unaware	1	Limited	1	Discouraging	1	Dictates to SP, TC *or* Absent	4

APPENDIX D: COGNITIVE INTERVIEW PROTOCOL

Introduction

Thank you for taking time today to speak with me about my survey. My dissertation has two main focuses in its research goals. My first goal is to develop a scale that uses a scenario-based item format, and the other is to learn more about the role of the university supervisor in teacher education.

Today, I am going to show you nine different scale items that use this scenario-item format. I am going to ask you to read and answer each item. I will then ask you about your thought process as you answered these items, and whether and words/phrases were confusing, or what images/memories came to mind as you read the scenario. There are no right or wrong answers to these items, as they are your personal thoughts and perceptions of your supervision experience.

To understand this conversation at its highest level, I am going to record the audio from our Zoom call today.

Are there any questions that you have before we get started?

Item 1

This is the first scenario question. I am going to display it on the PowerPoint, and I will give you some time to read it fully. Take as much time as you need. When you are finished reading, answer the question.

After individual has indicated they have read the scenario and answered the question.

Ok, now that you have finished, how did you answer the question?

How did this description compare to your supervisor?

Can you explain to me what you were thinking about as you read through this scenario?

Why did you pick the response option that you selected?

While reading the scenario, were there any words or phrases that were confusing?

Repeat for remaining eight items

During these next eight items, I will prompt the respondent to also reflect on the relationship between that scenario and previous ones.

Can you explain to me any similarities or differences you observed while reading this scenario, compared to other scenarios?

Conclusion

Thank you again for taking the time today to talk about my dissertation. I greatly appreciate it!