RASCH ANALYSIS OF A RATING SCALE FOR GIFTED AND TALENTED IDENTIFICATION

by

David Alan Nelson

A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
Doctor of Philosophy
Education

Committee:

_____ Chair

_____

_____

_____

_____ Program Director

_____ Dean, College of Education and Human Development

Date: _____ Fall Semester 2014
George Mason University
Fairfax, VA

Rasch Analysis of a Rating Scale for Gifted and Talented Identification

A Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

by

David Alan Nelson
Master of Science
University of Maryland, 2009
Bachelor of Education
University of Toledo, 2000

Director: Dr. Erin Peters-Burton, Associate Professor
College of Education and Human Development

Fall Semester 2014
George Mason University
Fairfax, VA

DEDICATION

I would like to dedicate this work to Rorey, who had to eat a lot of frozen meals, listened patiently to my explanations of an awful lot of item-person maps, item characteristic curves and test information functions, and tolerantly read draft after draft while I completed this dissertation. He took all of it in stride, and I appreciate that he did.

I would also like to dedicate this work to Brittany Hawkins, a former student of mine at Northern High School. Although she was taken from us way too soon, she inspires me every day and will continue to as I move forward to new experiences. She may never have known that I learned more from her than she ever could have from me.

ACKNOWLEDGEMENTS

I would like to acknowledge several George Mason University faculty whose guidance and support made both the program at Mason and the dissertation fulfilling experiences rather than tasks that needed to be completed. Dr. Erin E. Peters-Burton, my dissertation committee chair, for her valuable guidance in selecting a topic, feedback throughout the process and words of support that provided me the confidence that I was moving in the right direction. Committee member Dr. Anastasia P. Samaras for encouraging me to not get frustrated and keeping me calm when I couldn't find my "dot" right away–she knew I would find it long before I believed it! Committee member Dr. Robert G. Smith especially for providing opportunities in his courses to learn how to navigate the process of making the research-to-practice shift so that what we do can be used by those that put it into action with students. Committee member Dr. Ellen Rowe for working with me closely to find a topic on which to focus and helping me to work through the details of my first first-author presentation in the field–her guidance and feedback made it a much less anxious process. And, of course, Dr. Dimiter M. Dimitrov for not only introducing me to item-response theory and modern measurement methods but also for making them meaningful and, as he once said, "just a little bit sexy."

I would also like to acknowledge the great people at Mason who were not tied as closely to my dissertation but were essential in making sure I got here. Dr. Gary Galluzzo, Director of the Ph.D. in Education Program while I attended Mason–his leadership and vision surely is a large part of why the program was such an intellectually and personally rewarding experience. Joan Stahle, the Program Manager for the Ph.D. in Education Program at Mason–I barely had to think of the questions I had about how to navigate the administrative side of the program because Joan always ensured I had the answers I needed first. The faculty and staff of CEHD are amazing folks with whom I am proud to have had the opportunity to work!

I would also like to thank Dr. Joseph Renzulli and Dr. Del Siegle from the University of Connecticut for providing the data from their early validation studies for my use in this work.

And, on a personal level, I would especially like to acknowledge and thank Yovonda Ingram Kolo, who has helped me to grow professionally and listened to more about Rasch modeling than she ever cared to. I know that Yovonda's encouragement and support will not end with this dissertation–and I am honored to have her as a great friend.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND SYMBOLS

Chi-square statistic ..... χ^2
Comparative fit index ..... CFI
Confirmatory factor analysis ..... CFA
Degrees of freedom ..... *df*
Delta, item difficulty ..... δ
Differential item functioning ..... DIF
Differentiated Model of Giftedness and Talent ..... DMGT
Mean-square ..... MnSq
National Association for Gifted Children ..... NAGC
Principal component analysis ..... PCA
Rating scale model ..... RSM
Root mean-square approximation ..... RMSEA
Scales for Rating the Behavioral Characteristics of Superior Students ..... SRBCSS
Standardized *Z*-statistic ..... Zstd
Tau, threshold parameter ..... τ
Theta, ability estimate ..... θ
Tucker Lewis fit index ..... TFI

ABSTRACT

RASCH ANALYSIS OF A RATING SCALE FOR GIFTED AND TALENTED IDENTIFICATION

David Alan Nelson, Ph.D.

George Mason University, 2014

Dissertation Director: Dr. Erin Peters-Burton

A paradigm shift toward a talent development model of providing services in gifted education has transformed the traditional IQ-based notion of giftedness (Dai & Chen, 2013). The conceptualization of giftedness plays an important role in the development of programs for gifted and talented students, and at its core the conceptualization of giftedness plays a central role in the development of instruments for identification, including rating scales completed by teachers and other school-based staff (Borland, 2003; Robinson, 2009). This study used Rasch measurement analyses to evaluate the evidence of validity, characteristics of reliability, item selection, category structure and differential item functioning of the Scales for Rating the Behavioral Characteristics of Superior Students, 3rd Edition ([SRBCSS-III] Renzulli et al., 2013) using data from the SRBCSS authors' validation studies and data from Grade 3 and Grade 4 operational administrations in a Mid-Atlantic county school district. The Rasch rating scale model showed evidence for the substantive, structural and content validity

aspects of construct validity in both the original validation studies and the operational administrations. Strong item hierarchies, content representativeness and data reliability were shown in the analyses. Additionally, data from the operational administrations showed minimal differential item functioning for groups analyzed by sex, race/ethnicity and economic status. The study highlighted the strength of the psychometric properties of the scales while offering suggestions for improvement and future study in the context of additional aspects of validity.

Keywords: Rasch, validity, rating scales, SRBCSS, Renzulli, talent development

CHAPTER ONE: PURPOSE AND RESEARCH QUESTIONS

Contemporary conceptualizations of giftedness since the early 1990s reflect the view that giftedness is a socially-constructed, dynamic, fluid concept and not a stable and permanent category into which a small percentage of students fall (Borland, 2009; Callahan, Hunsaker, Adams, Moore, & Bland, 1995; Subotnik, 2003). Owing to the paradigm shift (Dai & Chen, 2013) toward talent development and the push away from a narrow IQ-based perspective of identifying for gifted education (Borland, 2003; Robinson, 2009), the ways and methods of identifying students for services in the areas of gifted and talented education have expanded.

A recent special issue of *Psychoeducational Assessment* (Pfeiffer, 2012b), highlighted a number of the perspectives on methods of identification, which include models rooted in response-to-intervention strategies, the use of nonverbal ability instruments, brief intelligence measures as part of a multiple methods approach, identifying through domain-specific instruments, and the use of wider examinations such as talent search and providing enrichment to all students. Indeed, the methods of identification are changing in response to the changing concept of what makes giftedness and how educators should be involved in nurturing it (Pfeiffer, 2012a).

Earlier, Feldhusen (2003) noted the particular role that teachers play in identifying students for services in the talent development programs–especially through the use of rating scales. Although this type of identification tool is not new, its use is growing because of the view that achievement and cognitive testing do not capture the aspects of students served in programs built on talent development models (Brown et al., 2005; Renzulli, 2012). Given this literature, areas of question arise that are salient in the context of the current study.

Statement of the Problem

Several researchers have evaluated the valid interpretation of scores using similar instruments, their reliability and psychometric characteristics (Jarosewich, 2002; Jarosewich, Pfeiffer, & Morris, 2002; Pfeiffer, Petscher, & Jarosewich, 2007). For the scales used in the current study, Scales for Rating the Behavioral Characteristics of Superior Students, 3rd Edition ([SRBCSS-III] Renzulli et al., 2013), an extensive author-conducted validation study is supplied as part of the background provided with the scales (Renzulli & Smith, 2010). Concerns presented in these have not been viewed in the context of modern measurement methods, which the current study will evaluate. In particular, four areas are addressed in the literature as areas in instrument evaluation needing researchers' attention. These are presented in the problem statement section that follows. Each will be developed in the context of results in the Discussion and Conclusion section later.

Construct validity. Jarosewich, et al. (2002) indicated a weak foundation of construct validity for the SRBCSS-III as well as several other scales used in the

identification process. This study evaluates evidence for construct validity in terms of Rasch analysis, which finds support for content, structural, substantive and generalizability validity using several metrics (Bond & Fox, 2001, 2007). Identification of item hierarchy on a scale supports content validity. Items must offer a range of difficulty along the continuum of the latent trait in order to represent a well-defined variable (Smith, 2003). Item reliability can be used to support the representativeness of the range of items, as well, and indicate items targeted for deletion or retention in a pilot study (Linacre, 2014a). Item reliability in the Rasch model is the extent to which the items on an instrument can be precisely located along a continuous latent variable. Classical measures of internal consistency such as Cronbach's alpha assume a continuous scale, which may not be the case with categorical data such as that on the SRBCSS-III (Bond & Fox; Linacre). Moreover, the utility of Cronbach's alpha is widely debated in many circumstances (Sijtsma, 2009). Importantly, differential item functioning has not been widely-studied in the application of rating scales for the SRBCSS-III (Renzulli & Smith, 2010), which is essential to ensuring the generalizability validity of inferences made using the scales. Moreover, the construct validity argument can build evidence that the conceptualizations of giftedness as talent development that appear in Chapter 2 form a well-defined description of current views.

Factor structure. Especially important in the current study is the use of Rasch in evaluating factor structure. Validation of the SRBCSS-III utilized confirmatory factor analysis, which often creates illusory factors, and performances at different levels can often present as different factors; Rasch principal component analysis can eliminate this

problem through principal component analysis of residuals (Linacre, 1998). Traditional confirmatory factor analysis results in the construction of variables, while Rasch principal component analysis identifies and explains variance after the contribution of the measure is removed (Linacre, 2014a). Kreiner and Christensen (2011) offer an argument for the advantages of Rasch analysis of factor structure under certain conditions, including those where authors have moderate confidence in the items, which is more often the case where instruments are built upon ill-defined constructs.

Scale functioning. In order to be useful and support reliable, valid measurement, rating categories must be substantively different and meaningful for respondents (Linacre, 2002). Rating categories for polytomous items such as those on the SRBCSS are evaluated as individual dichotomies between category k and the adjacent category $k - 1$. Essential to the functioning of such a scale to produce a measure is the advancement monotonically of the categories at higher levels of trait (Dimitrov, 2012). Threshold or category disordering is not evaluated in typical evaluations of scales, while with Rasch modeling such diagnoses can be made to evaluate the effectiveness of rating scale structure. This provides an instrument's author insight into the substantive meaning interpreted by respondents to ensure support for its use rather than relying on conjectural or anecdotal information to make decisions about category labels or category number.

Fundamental measurement. In order to be used in such a way as described in much of the literature (Pfeiffer, 2007; Renzulli & Smith, 2010), scores must be able to be meaningfully summed if the goal is to do so in order to arrive at overall scores. However, raw scores on such instruments do not have the measurement capacity to be summed, as

they are categorical, ordinal data. Rasch modeling provides the property of fundamental measurement to such data in the form of additive conjoint measurement by transforming the scores into interval data with items and persons on the same scale (Bond & Fox, 2007). Thus, if the data conform to the Rasch model, such fundamental measurement can support the use of the rating scales in a number of ways.

Data and Participants

The current study uses data from both the SRBCSS-III authors' validation study as well as an operational administration in a public school district in a mid-Atlantic state. The first data set, the SRBCSS-III authors' validation study data set, includes the field test item-level response data gathered as part of the SRBCSS-III domain scales validation study performed when the scales were added to the SRBCSS, 2nd Edition in 2010. These data include the item-level response data from a 73-item field test, which was ultimately reduced to a 30-item set from which four domain scales were constructed–reading, mathematics, science and technology–for inclusion in the current version of the SRBCSS-III (Renzulli & Smith, 2010).

The second data set was collected as part of the annual identification process for placement into a talent development setting for Grade 3 and Grade 4 students. In the district, this operational administration of the rating scales are performed as a census administration for all Grade 3 and Grade 4 students, as placement into a talent development setting first occurs as students enter into Grade 4.

Purpose of the Study & Research Questions

The purpose of this study is to use Rasch analysis to evaluate the validity, characteristics of reliability, item selection, category structure, and differential item functioning of the SRBCSS-III, which are commonly used in identifying students for placement in gifted programs. Ostini and Nering (2006) suggest that at the most fundamental level mathematical models provide a means to quantify phenomena. Importantly, the instruments by which phenomena are measured should provide evidence for aspects of validity, provide reliable data scores, and be efficient to use (Dimitrov, 2012; Linacre, 2002). The primary purpose of this study is to apply Rasch measurement to six scales of the SRBCSS-III to answer the following research questions.

Research Question 1: Does Rasch analysis confirm the dimensionality and evidence of well-functioning retained items on the domain scales (reading, mathematics, science and technology) added to the SRBCSS-III?

Research Question 2: Does Rasch analysis provide evidence for reliability and validity for the domain scales (reading, mathematics, science, and technology) and the learning characteristics and motivation scales on the SRBCSS-III?

Research Question 3: Does Rasch analysis show optimal category structure for the SRBCSS-III?

Research Question 4: Is there evidence of differential item functioning for subgroups of students on selected scales of the SRBCSS-III?

Definitions

(Bond & Fox, 2001, 2007)

Construct: A single latent trait, attribute or dimension underlying a set of items.

Construct validity: An argument that items on an instrument are true operationalizations of the latent trait, attribute or dimension being measured.

Content validity: An argument for validity that includes content relevance and representativeness and technical quality.

Differential item functioning: The occurrence of item difficulty being different for compared groups after adjusting for their similar overall abilities.

Error estimate: The difference between an observed and expected response.

Fit: The degree of match between pattern of observed responses and the modeled expectations.

Fit statistic: An index expressing the match between observed patterns and modeled expectations.

Fundamental measurement: Measurement resulting in additivity illustrated by physical concatenation.

Infit statistics: Expresses the degree to which on-target observations fit modeled expectations.

Interval scale: A measurement scale in which the unit of measurement is maintained throughout the scale.

Item separation: An estimate of the spread of items along the measurement continuum for a variable.

Log odds ratio: The logistic transformation of the odds ratio,
ln [probability of success / (1 – probability of success)].

Logit: Log odds ratio contraction.

Measure: The Rasch estimate of item difficulty or person trait level.

Operational administration: An administration of an instrument that will be used for making placement decisions.

Ordinal scale: A measurement scale that rank orders with no magnitude difference between ranks specified.

Outfit statistics: Expresses the degree to which off-target observations fit modeled expectations.

Point-measure correlation: The correlation between the observations in the data and the Rasch measure.

Rasch rating scale model: A version of the family of Rasch models that requires the same number of categorical choices in each item and applies one set of threshold values to all items.

Residual: A value representing the difference between the Rasch model's expectations and actual observations.

Specific objectivity: Results when measurement is independent of the sample of items or the sample of persons being measured. Rasch measurement has specific objectivity.

Stochastic: Illustrated by models that express probabilistic expectations of performance rather than exact expectations (deterministic).

Threshold: The level at which failure to endorse in a category results in the likelihood of endorsement in the next category.

Unidimensionality: A concept of measurement that one attribute be measured at a time.

Validity: Evidence gathered that supports inferences made from responses to measurement of a construct.

CHAPTER TWO: REVIEW OF THE LITERATURE

The purpose of this study is to use Rasch analysis to evaluate the validity, characteristics of reliability, item selection, category structure and differential item functioning of a rating scale instrument, the Scales for Rating the Behavioral Characteristics of Superior Students, 3rd Edition (Renzulli et al., 2013), which is commonly used in identifying students in Grade 2 through Grade 12 for placement in programs for the gifted or for talent development. In the literature review five major areas are discussed to provide context and background to the instrument, methodology and conclusions presented later. First, conceptions of giftedness are presented, which are followed by a review of models of talent development as frameworks for conceptualizing new ideas surrounding gifted education. Next, the previous validation study of some subscales on the SRBCSS-III are discussed. Finally, a presentation of the underlying considerations for scale development, including validity, unidimensionality, reliability and the use of rating scale categories precedes the review of the Rasch model and its characteristics and meaning that concludes the literature review.

Gifted and Talented Research

The notion of giftedness continues today to be considered a sometimes socially-constructed paradigm (Borland, 2003) and other times simply innate ability (Jensen, 1997). At any rate, it is a notion on which a common, cohesive definition or

conceptualization has yet to be agreed (Ambrose, Van Tassel-Baska, Coleman, & Cross, 2010). Dai and Chen (2013) suggested that a shift in what they term the "chaos" (p. 151) of an apparently incoherent body of knowledge is needed to clarify a meaningful relationship between the research in the field of gifted education with the practice of understanding educating the gifted. Specifically, Dai and Chen related such an understanding as having several characteristics: an assumption of the nature of gifted students and their educational needs, a purpose for the services provided in the educational setting, a clear identification process that is consistent with the assumption and purpose, and an articulation of the practices that can be used to successfully accomplish the purpose. They noted, however, that the fundamental problem with the conceptualization of giftedness is not a lack of definitions or attempts at identifying parameters of giftedness. Instead, it arises from the astoundingly large number of definitions and often competing theories. Importantly, the conceptualization of giftedness plays an important role in the development of instruments for identification–including the SRBCSS-III–and should be well-developed in theory and practice.

The next section highlights literature on how giftedness is conceptualized by both researchers and practitioners. Notably absent are conceptualizations based mainly on achievement test scores or IQ scores. Such metrics do not form the foundation for the development, administration or interpretation of the SRBCSS-III (Renzulli & Smith, 2010), as they should be additional components to a comprehensive identification program using such scales. Thus, they are not relevant in the current context.

Conceptual foundations of giftedness. The conceptual foundation of giftedness has fundamentally changed over the forty years since the Marland Report (Marland, 1972), which promulgated the first federal-level definition of giftedness and influenced Renzulli (1978) in his revolutionary (at the time) conception of giftedness in schools (Renzulli, 2014, personal communication). The conception of giftedness later experienced a second major shift in the early 1990s (Callahan et al., 1995) that continues today (Renzulli, 2012). Instead of remaining solely within the domain of intellect (Callahan, et al.) the conception has followed along at least two strands beyond traditional definitions. Along one strand is the conceptualization that giftedness (and, as discussed later, the opportunities for talent development) includes domains inside and outside of school that extend beyond traditional academics such as math, reading, and science. Along another is the conception that giftedness is defined in terms of specific behaviors rather than intellectual ability scores alone (Robinson, Shore, & Enersen, 2007). Indeed, in many modern representations, the strands readily intertwine. This provides a blended understanding of giftedness that allows for greater opportunities for students that exhibit the indicators widely viewed as those found in gifted students or those with recognized potential (Olszewski-Kubilius & Lee, 2004).

In the remainder of this section, several perspectives in the extant literature relating to the conceptualization of giftedness are presented. Commonalities between these are reflected in the items selected for the SRBCSS-III, which will be explicated in the Discussion and Conclusions section later. Such commonalities build validity evidence for the use of rating scales as part of a comprehensive identification program in schools.

In an attempt to synthesize modern thinking surrounding the conceptualization of gifted education, Cohen (2005) performed a large-scale analysis of the abstracts in the Conceptual Foundations strand sessions of the National Association for Gifted Children (NAGC) conferences from 1989 to 2004. In her analysis, Cohen noted the emergence of nine themes underlying the conceptualization and definition of gifted education, which are shown in Table 1. The varied nature of the strand presentations illustrate the number of directions in which the researcher or practitioner of gifted education must look in hopes of finding a consensus on the nature or relationships on which to focus in defining giftedness.

Table 1

Strand presentation themes, NAGC Conceptual Foundations Division, 1989-2004

Conceptions of giftedness	Individual characteristics (affective, intellectual, and thinking)
	Interaction with the environment
	Context (social, political, and diversity)
	Practice
	Issues and directions
Definitions, meanings & constructs of giftedness	Perspectives (changing and conflicting)
	Focus on intelligence, creativity, or talent
	Individual differences
	Definitions of domains

Note. Adapted from Cohen, 2005.

Schroth and Helfer's (2009) review of the literature presented an alternative synthesis of the conceptualization of giftedness in grouping ideas with respect to the behaviors of students or the purported purposes for offering gifted services in school to a sub-population of students or the census population. For example, giftedness can be defined in terms of superior performance and ability, appearance of potential or as a mechanism of social justice and equity in a school. Schroth and Helfer took the position that there is no need to adopt one or the other of these as if they are mutually exclusive, as each captures an aspect of philosophies around which programs can be built. Instead, the essential piece is consistency in practice.

Subotnik, Olszewski-Kubilius and Worrell (2011) proposed a definition that incorporates many of the widely-held beliefs surrounding the conceptualization of giftedness that contains hallmarks of the two large-scale syntheses conducted prior and discussed above. They suggested that giftedness: reflects the values of society; manifests in actual outcomes; is domain specific; is the result of psychological, social and physiological factors; and is relative in nature to ordinary and extraordinary ability on a continuum.

Like Csikszentmihalyi (1997) and Gagné (1995, 2009), Subotnik et al. (2011) distinguished purposefully between giftedness and talent, which is not always the case and often further confounds the conceptualization in the literature (Gagné's theory will be further developed in the next section). Subotnik et al. asserted in their conceptualization that ability is necessary for giftedness, but it is not sufficient for the development of a talent in an area of giftedness. Interest, commitment, appropriate teaching and coaching

are essential to full development of a talent. They further suggested that developmental periods vary across different domains and the emergence of new domains give rise to the possibility of identification and emergence of new aspects of giftedness.

The argument presented in the Subotnik et al. (2011) article attempts to bring clarity to the wide variety of conceptualizations of the essential components needed to explain or observe giftedness. In fact, the authors presented a coherent set of contributors to giftedness and the ways in which educators should respond to giftedness that are useful in later discussing the development of rating scales for identification. These are shown in Table 2. These conceptual aspects of giftedness are integrated into many of the theoretical models of giftedness, which will be discussed in the next section. These conceptual aspects will be discussed as informing the development of the SRBCSS-III.

Table 2

Contributing and Encouraging Characteristics of Giftedness

Area	Characteristics	
Contributing	Ability	Interest
	Creativity	Passion
	Motivation	Opportunity
	Personality	Chance
	Emotions	Parents
Encouraging	Enrichment	Psychosocial coaching
	Acceleration	Selective placement

Note. Adapted from Subotnik et al., 2011.

Important to the development of giftedness and gifted programs in schools are the conceptions of teachers in classrooms, as well. Moon and Brighton (2008) conducted a survey research project to evaluate primary teachers' conceptions of giftedness and talent development through their work with the National Research Center on Giftedness and Talent. Among the characteristics identified by more than 50% of teachers in the Moon and Brighton study as very easy to imagine about a gifted student are that he or she

- easily transfers learning to other subjects or real-life situations;
- tries to understand the 'whys' and 'hows;'
- has a large store of knowledge;
- has an active imagination;
- likes to make 3-D structures;
- completes assignments faster than peers;
- can adapt strategies;
- carries on conversations with adults;
- has unusual interests for their age; and
- can carry out multiple verbal instructions.

Again, as with Subotnik et al.'s (2011) definition, these echo the types of characteristics that are often found on teacher rating instruments for identification.

Talent development models. As mentioned in the previous section, Ambrose et al. (2010) and Dai and Chen (2013) found an apparent disconnect between research and practice in gifted education. In the context of the current study, an instrument purporting to measure the characteristics of gifted students should be firmly rooted in the theory of

giftedness and have a model as its framework (Jarosewich, 2002; Jarosewich et al., 2002; Pfeiffer et al., 2007). Today, many of the existing instruments have been constructed around the talent development model.

Talent development models arose as a reaction to several factors, one of which was the 1972 so-called Marland Report, *Education of the Gifted and Talented*:

> Students, children, or youth who give evidence of high achievement capability in areas such as intellectual, creative, artistic, or leadership capacity, or in specific academic fields, and who need services and activities not ordinarily provided by the school in order to fully develop those capabilities. (National Association for Gifted Children, n.d.)

This definition, later incorporated into the Elementary and Secondary Education Act and its reauthorizations (U.S. Department of Education, 2014), explicitly alluded to non-academic domains, which had largely been excluded from the conversation prior to the report's publication (although not completely; e.g., Witty, 1958). After the Marland Report, successive researchers, including Tannenbaum, Renzulli, Gardner, Bloom, Sternberg and Gagné, provided the impetus to supplement the IQ/gifted child philosophy with the talent development model.

However, the outcome of the talent development model is not to supplant the fact that ability (e.g., IQ greater than 130) can be a "strong and real" (p. 13) sign of giftedness as argued by Delisle (2003). Instead, the talent development model has grown alongside more traditional views of giftedness to nurture potential by providing opportunities for all students to receive an enriched early childhood education, reinforce commitment to tasks,

and guide individuals to explore interests, which idealizes an outcome of adulthood expertise in a domain (Subotnik, 2003).

This section reviews the literature on four models of giftedness grounded as talent development models, as these are the models that the SRBCSS-III is designed to support (Renzulli & Smith, 2010). First, an overview of the paradigm of talent development is presented; then, an exposition on four widely-cited, well-developed, and theoretically-based talent development models follows. Such models build the foundation supporting the validity of the interpretations on rating instruments such as the SRBCSS-III (Pfeiffer, et al., 2007; Renzulli, 1978; Renzulli, personal communication, 2014).

The talent development paradigm. Olszewski-Kubilius (2009) stated the paradigm shift toward talent development as a model of gifted education gained momentum in the mid-1980s, which was the time of Bloom's (1985) and Gardner's (1985) seminal works, both of which affirmed the work of Renzulli (1978), in particular, in discussing the nature of domain-specific talent leading to a realized giftedness. Moreover, the contextual dependence of talent development on family, school and community is especially prominent in the work of all three authors. Olszewski-Kubilius reminded us that Bronfenbrenner's ecological theory (1979) asserts just such a dependence: children exist within a series of settings or contexts that influence development and that optimal conditions within each context increase the likelihood of drawing on a child's potential.

Dai and Chen (2013) framed their discussion of the talent development paradigm in terms of the essential characteristics discussed earlier: assumptions (‘*What*?’), purpose (‘*Why*?’), target (‘*Who*?’), and strategies (‘*How*?’). The answers to these questions, shown in Table 3, illustrate the movement toward a developmental, contextualized model reflecting diverse representations of giftedness across a broader range of students and represent the view of talent development across a range of theoretical perspectives.

Table 3

Characteristics of Dai and Chen’s Talent Development Model

Assumption ‘*What*?’	Talent development is a malleable set of cognitive and non-cognitive developmental capabilities; it involves the evolving domain-specific nature of talent, motivation, opportunity, differentiated trajectories and social support.
Purpose ‘*Why*?’	Talent development is to cultivate a range of strengths and interests; teachers decide on timing and trajectory of specialization and the degree of coaching and mentoring; stresses unique contributions.
Target ‘*Who*?’	Heterogeneous groups of students; identification involves cognitive and non-cognitive domains; students might self-select into clubs, organizations or activities; broad enrichment provided to all.
Strategies ‘*How*?’	Providing interest-based experiences, real-world and authentic tasks; involvement with experts; long-term involvement; extension beyond the classroom and beyond academic coursework.

Tannenbaum's talent development model. Tannenbaum's (1983, 1997, 2003) model clearly illustrates the developmental nature of talent from childhood to adulthood, and he asserted that students able "…to make the most of enriching experiences…" (1997, p. 39) can grow up to be gifted. Tannenbaum clearly distinguished giftedness from talent development in that it is the development of talent over time that leads to the expression of giftedness as adults (Subotnik et al., 2011).

The model supposes that there are two kinds of gifted people: producers and performers (Tannenbaum, 1983, 1997). In both *creative* and *proficient* ways, producers are those that exhibit talent in producing *thoughts* and *tangibles*, while performers demonstrate talent in the areas of *artistry* and *human services*. Typical people that exemplify the talents developed by producers and performers are shown in Table 4; indeed, characteristics of those considered experts in such fields should be echoed in instruments purporting to identify students as gifted through the lens of talent development.

Within the model, Tannenbaum (1983, 1997) theorized elements that contribute to the development of gifted behavior: superior mental intellect, distinctive mental aptitudes, a supportive set of nonintellective characteristics, a challenging and supportive environment, and chance. Importantly for the identification of talent in students is Tannenbaum's assertion that giftedness is not plausible if any of the five elements is absent. Although the "theoretical [physicist] requires higher general ability and fewer interpersonal skills…[t]he five factors interact in different ways for separate talent domains, *all* are represented in some way in every form of giftedness" (1997, p. 30).

Table 4

Characteristics of Tannenbaum's Producers and Performers

Producers	Philosophers, writers, composers, scientists and historians (creative thought)
	Mathematicians, computer programmers, and editors (proficient thought)
	Inventors, sculptors, artists, architects and engineers (creative tangible)
	Machinists, masons and technicians (proficient tangible)
Performers	Fine arts performers who perform interpretively to their interpretation; debaters (creative artistry)
	Fine arts performers who perform faithfully to an author's interpretation (proficient artistry)
	Teachers, political leaders, clinical workers, researchers (creative human services)
	Medical doctors, administrators, psychiatrists , CEOs (proficient human services)

Note. Adapted from Tannenbaum, 1997.

In summary, in his talent development model Tannenbaum (1983) fundamentally suggested that aptitudes and abilities are essential characteristics of giftedness that can be developed in nurturing, supportive environments. However, a successful transition to giftedness is only possible when the aforementioned are accompanied by motivation, determination, and perseverance.

Sternberg's developing expertise model. Growing out of his triarchic theory of intelligence (1985, 1995), Sternberg's developing expertise model (1999, 2002, 2001) characterizes how three areas of intelligence (analytical, creative and practical) can be viewed as aspects of a development process resulting in a high level of mastery in one or more domains.

In a study (Sternberg, 1995) in which construct validation was being sought for the triarchic theory of intelligence, gifted high school students were tested for their analytical, creative and practical abilities. The researchers found that different patterns of ability were reflected for the three aspects of intelligence, which resulted in a low general intelligence factor. They concluded that the three aspects of ability were, indeed, separate forms of developing expertise rather than a singular or unitary general intelligence factor. The implications of this research suggests a talent development model in gifted education as well in the broader range of classrooms in schools (Sternberg, 1999, 2001).

The model arising from Sternberg's research (1999, 2001, 2002) has five elements, which Sternberg acknowledged as not exhaustive in the development of expertise: metacognitive skills, learning skills, thinking skills, knowledge and motivation. Students identified as gifted excel in one or more of the interactive elements. Importantly, the elements are generally domain-specific; that is, development of expertise in one area does not necessarily result in expertise in another. And, the elements can influence each other directly and indirectly, as when metacognitive skills influence the development of thinking skills.

Central to the model as a model of talent development is, in fact, the interaction of elements (Sternberg, 2001). Sternberg provided examples of motivation driving metacognitive skills, which then leads to increased learning and thinking skills that finally returns to further reinforce the development of metacognitive skills. Procedural and content knowledge developed through the process results in the ability to use all of the skills more effectively later. As such, the process of talent development is a feedback

loop that continues dynamically toward developing expertise and exhibiting characteristics of giftedness. A summary of the elements with explanatory descriptions of their features in practice is provided in Table 5.

Table 5

Elements and Features of Sternberg's Developing Expertise Model

Element	Description	Features
Metacognitive Skills	Understanding and controlling one's cognition	Problem recognition, definition, and representation; formulation of strategies; allocating resources; monitoring and evaluating problem-solving
Learning Skills	Components of knowledge acquisition	Encoding; distinguishing relevant and irrelevant information; synthesizing information; relating new and old information
Thinking Skills	Performance components	Critical (analytical) thinking, including analyzing, critiquing, judging, evaluating and assessing; creative thinking skills, including creating, discovering, imagining, supposing and hypothesizing; practical thinking skills, including applying, utilizing and practicing
Knowledge	Knowing	Declarative knowledge, such as facts, concepts, principles, and laws; procedural knowledge such as procedures and strategies
Motivation	Drive	Achievement motivation, relating to seeking (moderate) challenges and risks leading to improvement in some domain; competence, or self-efficacy motivation, relating to believing in one's ability; growth motivation, relating to wanting to develop one's intellectual skills

Renzulli's three-ring conception of giftedness. Renzulli (1978) proposed the three-ring conception of giftedness at a time when the definitions of giftedness were diverging in the field in the years following the release of the Marland Report. Renzulli cited three problems with the definition of giftedness provided by the U.S. Office of Education (at that time the U.S. Department of Education had not yet been established): the definition was often misinterpreted and misused by practitioners, the categories were non-parallel in their construction, and nonintellective factors were not included.

In developing a model around which giftedness can be defined, Renzulli (1998) noted four criteria that must be met.

1. The model must be based on research relating to gifted children.
2. The model must provide guidance for instruments that can be used and procedures that can be followed to support the model.
3. The model must give direction to the programming, training and evaluation components that can be used to support the model.
4. The model must be researchable itself in any attempt to validate the definitions used within it.

Framing a model in such a way, Renzulli asserted, ensures that a logical, defensible and purposeful relationship exists between the definition and the programming and identification processes used to support the model.

Renzulli (1986, 1998) suggested two kinds of giftedness that underlie the three-ring model: schoolhouse giftedness and creative-productive giftedness. Schoolhouse giftedness includes the typical classroom-type ability, and it is often the type that is most

valued in the classroom. Often, this manifests itself as high scores on cognitive assessments or achievement tests, although it would be detrimental to use these as the only measure. For example, IQ tests correlate only .40 to .60 with school grades, indicating that as much as 84% of the variance in school grades is not related to a child's IQ. Creative-productive giftedness relates to the aspects of activities that result in the development of products for an audience. These are the activities that were not captured by traditional assessment for gifted or talent development programs at the time Renzulli first put forth his three-ring conception of giftedness (Renzulli, 1978). Importantly, the model proposes that these two kinds of giftedness interact, and both should be encouraged through special programs and opportunities for students.

Fundamental to Renzulli's (1978, 1986, 1998; Subotnik et al., 2011) three-ring conception model is that giftedness is not a characteristic in a person. Instead, behaviors of those that have developed giftedness can be observed. And, when these are observed in children, then it is incumbent on educators to "provide young people with maximum opportunities for self-fulfillment through the development and expression of one or a combination of performance areas where superior potential may be present" (Renzulli, 1998, p. 109).

The three-ring conception (Renzulli, 1978, 1986, 1998; Renzulli & Smith, 2010; Renzulli et al., 2002) emphasizes the research that shows how clusters of behavioral characteristics that are possessed by those demonstrating giftedness (or, Sternberg's *expertise*) interact to explain giftedness. In the model, above-average ability (addressing schoolhouse giftedness), creativity (addressing creative-productive giftedness) and task

commitment (addressing the component missing from the U.S. Office of Education definition following the Marland Report) are the clusters that interact to move students toward developing talents into giftedness, as shown in Figure 1; it is the point of overlap among all three clusters where the truly gifted are located. It is important to note two clarifying points relating to the model before discussing each cluster (Renzulli, 1978, 1998).

1. Each cluster plays an important role in contributing to the display of gifted behaviors, and an emphasis on superior ability in one area at the expense of the other two is an error in the model's application.
2. No single cluster makes giftedness; instead, the interaction among the three equally-important clusters has been shown in research to be necessary for gifted accomplishment.

Above average ability. Two types of ability are considered in the three-ring conception (Renzulli, 1978, 1998). General ability refers to the capacity to process information, integrate experiences, and to engage in abstract thinking. Verbal and numerical reasoning, spatial relations, memory and fluency represent examples of general ability. These can be measured through aptitude and achievement tests. Specific ability refers to the capacity to acquire knowledge, skill or the ability to perform in specialized activities. Although they can sometimes be measured in the same way, specific abilities can be expressed in non-test situations. Applying the academic domains (chemistry and mathematics, for example), composing music, dance, photography and art are examples of specific abilities. Renzulli (1998) noted the extensive research base that suggests at

best a slight relationship between general ability and specific ability. As such, it remains desirable to extend the criteria of talent to the areas of specific ability.

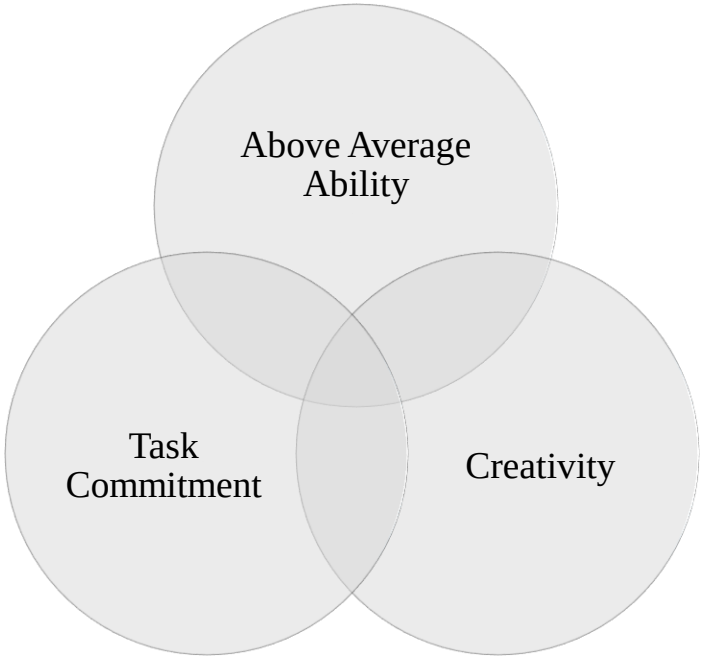

Figure 1. Renzulli's three-ring conception of giftedness. From "Reexamining the Role of Gifted Education and Talent Development for the 21st Century: A Four-Part Theoretical Approach," by J. S. Renzulli, 2012, *Gifted Child Quarterly, 56*, p. 152. Copyright 2012 by National Association for Gifted Children. Reprinted with permission.

Creativity. Creativity relates to the capability to set aside established conventions, try novel procedures, plan, construct resourcefully, and think in divergent ways (Renzulli, 1978, 1998). Importantly, even today this area has an ill-defined research base around

which to build as strong of a case as can be with ability and motivation in terms of leading to giftedness through talent development.

Task commitment. Renzulli (1978, 1998) explained task commitment as a focused form of motivation. Renzulli asserted that a necessity of those exhibiting giftedness is their total involvement in a specific problem for an extended period of time, or task commitment. Defined as representing a commitment to expend energy on a particular problem or performance, task commitment is often described by terms such as perseverance, hard work, dedication, and self-confidence. Reinforcing the role of task commitment in reaching expertise through talent development, Renzulli recalled (1998) that even Terman acknowledged the less-than-perfect correlation between achievement and intellect. While Renzulli reminded us that Terman never adopted a view of a nonintellective conception of giftedness, Terman referenced the importance of persistence, integration toward goals, self-confidence and drive to achieve in his works.

Gagné's differentiated model of giftedness and talent. Gagné's differentiated model of giftedness and talent ([DMGT] 1985, 1999, 2004, 2009) explicitly distinguishes between giftedness and talent. Like Tannenbaum (1983), Gagné made a distinction between a child's emerging talent and an adult's expression of giftedness, but acknowledged three overlapping characteristics: (a) both refer to human abilities, (b) both target individuals who differ from the average, and (c) both involve those with outstanding behavioral attributes (2009).

Gagné (1985, 1999, 2009) put forth a definition of the talent development process, which is the third of three components in his DMGT. The definition entails the

process of talent development through which gifts become talents. And, mediating the talent development process are intrapersonal catalysts and environmental catalysts.

Gagné (2009) identified gifts (*G*) as the natural abilities such as intellectual abilities, creative abilities, and social abilities. These are observed in the daily tasks related to schooling. Among the abilities in the domain of *G* are verbal, numerical, spatial, procedural and reasoning ability in the intellectual domain; imagination, originality, and problem-solving in the creativity domain; and social ease, tact, influence, persuasion, and leadership in the social domain. However, Gagné asserted that they are not innate abilities but that they instead develop over the life of a person–especially developing in children in their early years.

In the talent (*T*) component, Gagné (2009) identified the competencies associated with several fields in which talent is often seen. Especially significant in terms of providing support for the identification of students in the early years are the academic (language, math, science, humanities, and vocational), science and technology (engineering, medical, and social), and arts (creative, performance, and applied arts) fields.

As with other models of talent development, the DMGT has at its core the relationship between the abilities people have and the manifestations of those abilities as they develop into demonstrable talents (Gagné, 1985, 1999; Renzulli, 1978). The talent development component (*D*) of the DMGT involves the guided, structured and supported pursuit of excellence in a field. In DMGT 2.0 (2009), Gagné put forth a model by which this pursuit occurs, a model in which activities (access, content and format), measured

progress (stages and turning points) and investment (time and energy) work to encourage the transition from *G* to *T*. Figure 2 is an adaptation of DMGT's framework components as they relate to the conceptualization of giftedness in terms of the current study.

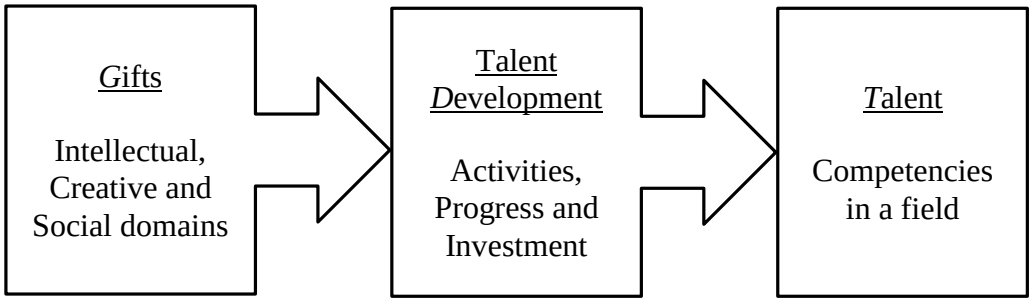

Figure 2. Modified process map of DMGT.

The SRBCSS-R and SRBCSS-III

In the preceding section, sets of characteristics that form several models for talent development have been identified, which begins to build the theoretical foundation for assessing and supporting construct validity (Messick, 1995a) for the SRBCSS-III. Of course, the four models discussed are not the only models in the talent development arena. Feldman's (1988) co-incidence model, Stanley's (Assouline & Lupkowski-Shoplik, 2012) talent search model, and Subotnik et al.'s (2011) mega-model are additional examples of models designed around the idea that developing talents provides trajectories toward giftedness. But, the four models reviewed above are particularly salient in that they represent a significant body of the research used in the development of the SRBCSS-III (Del Siegle and Joseph Renzulli, personal communication, 2014) and

describe well the characteristics and behaviors of gifted students. In the next section, the research conducted to develop the scales added to the SRBCSS-R and SRBCSS-III is reviewed.

Validity studies of the SRBCSS-R and SRBCSS-III. Validity and reliability studies for the motivation and learning characteristics scales were completed when the SRBCSS-R were released in 2002 (Renzulli et al., 2002). Similar studies for the reading, mathematics, science and technology scales were conducted when these scales were added later as part of the SRBCSS-III (Renzulli & Smith, 2010). Here, an overview of those studies is presented to provide additional context to the current study.

The domains of reading, mathematics, science and technology were defined by the behaviors observed by those talented in the domains as described in extant literature (Renzulli, Siegle, Reis, Gavin, & Systma Reed, 2009). Unfortunately, the authors did not clarify the theories, frameworks or resources referenced in selecting the original set of items for the content domains added in 2010. However, they do reference the emerging field of technology as difficult to conceptualize in terms of behaviors. Nonetheless, the domain is generally defined for the SRBCSS-III as related to computers and software, as well as communicating in the information technology field. Certainly, and as the SRBCSS-III authors noted, this domain will need continued refinement and its scale revised.

SRBCSS-R. Four changes were made to the original SRBCSS items as part of the revision published as the SRBCSS-R: compound items were removed, gender neutral pronouns were substituted for gendered pronouns, new items were added owing to the

interim research in giftedness occurring since the original scales were published, and syntax changes related to ensuring consistency were made (Renzulli et al., 2002). Of particular relevance to this study, a major change in the number of categories and the category labels was made in the revised edition. Renzulli et al. reported that teachers and specialists in gifted and talented education were dissatisfied with the original four categories available for each item. The users of the SRBCSS reported having difficulty perceiving the four category scale as an interval scale. Thus, a revised six category response scale was devised that remains in use today on the SRBCSS-III.

The construct validity of the SRBCSS-R was evaluated by a panel of 53 experts in the field of gifted and talented education. In a back-and-forth process of presenting items to the experts and removing and replacing those that did not meet minimum thresholds of domain placement agreement or strength of agreement within a domain between experts, the final SRBCSS-R was field tested with 54 items. Analysis of the SRBCSS-R is not related directly to the current study; readers are directed to Renzulli et al. (2002) for the details of the field test sampling methods. However, it is salient to note that the field test data on which construct validity was based were collected from teachers rating students identified as above average according to certain metrics or identified as above average in potential as identified by their teachers and not on a census sample of students with a wide range of ability.

After minor revisions to items based on the field tests, 71% of the variance was explained by a four-factor solution, which was consistent with the experts' category placement, lending support for construct validity and acceptance of a four domain

structure for the scales added to the SRBCSS-R (Renzulli et al., 2002). To provide support for the criterion-related validity of the SRBCSS-R, a number of teachers completed a supporting instrument on a small sample of students. The SRBCSS-R authors concluded that $R = .42$, $p < .001$, explaining 17.6% of the variance provided sufficient evidence for criterion-related validity of the SRBCSS-R.

Inter-rater reliability was established on a small sample of middle school students that were rated by two teachers: one teacher of mathematics and one teacher of language arts (Renzulli et al., 2002). A Pearson coefficient between teachers rating the same student of $r = .50$, $p < .01$and an intraclass correlation between the two rating groups of $r = .65$ provide moderate support for inter-rater reliability. However, the SRBCSS-R authors suppose that the moderate value is obtained owing to the different characteristics that would be observed by mathematics and language arts teachers.

SRBCSS-III. In 2010, the SRBCSS-R was updated to include four new scales that allow teachers to obtain ratings on students in four specific content domains: science, reading, mathematics, technology, and science (Renzulli et al., 2009). The authors' stated purpose was to encourage the identification of students in particular content areas that could lead to content-specific advanced academic programs, differentiated instruction or acceleration (Renzulli & Smith, 2010; Renzulli et al.).

As a first step in gathering support for construct validity, content area specialists in reading, mathematics, science and technology education reviewed the literature to identify the behavioral characteristics of gifted students in these areas. From literature reviews, a list of potential items was constructed for each content area, which was

evaluated by separate groups of professionals and by groups of general educators. The professionals indicated the strength of their agreement (1 = *highly appropriate*, 2 = *appropriate with editing*, or 3 = *not important/exclude*) to the appropriateness of how each item conceptually described the characteristics identified in the literature reviews, and each participant selected the items he or she considered the top 10 to 15 items to be used in the content area he or she was judging. In addition, the professionals indicated the most appropriate grade level for each item stem. Each proposed new subscale was evaluated by a minimum of 25 professionals. Items receiving strength ratings of two or three from 80% or more of the ratings were used in the pool of 73 field test items. The field test was constructed of all 73 items in a single instrument, which was completed by 187 teachers rating 726 students (Renzulli et al., 2009).

Confirmatory factor analyses (CFA) was performed on the field test data using Amos 4.0. First, separate CFAs were performed on each of the four domain scales. Once items were removed to obtain the best fit models for the individual scales, a final CFA was conducted on a model that included all four domain scales together. Model fit was evaluated using chi-square (χ^2), ratio of chi-square-to-degrees-of-freedom (χ^2/df) root mean-square error of approximation (RMSEA), and comparative fit (CFI) and Tucker Lewis fit (TLI) indices. (Renzulli et al., 2009).

SRBCSS-III results. Renzulli et al. (2009) reported reliability for each of the new scales added to the SRBCSS-R as Cronbach's alpha and reported CFA fit indices as discussed above. Results of CFA fit indices and reliability estimates for full scales are

provided in Table 6. Item-level descriptive information on retained items is provided in Appendix B.

Table 6

SRBCSS-III Domain Scales and Overall Model CFA Fit Indices and Internal Consistency Reliability Estimates, Best-Fitting Models

Scale	Items Retained (Initial No. Items)	RMSEA	χ^2/df	CFI	TLI	Cronbach's alpha
Reading	6 (9)	.052	2.92	.996	.993	.964
Mathematics[a]	11 (20)	.084	5.92	.978	.972	.977
Science	7 (24)	.074	4.80	.987	.981	.947
Technology	7 (20)	.060	3.25	.993	.990	.959
Combined Scales	30 (73)	.070	4.15	.945	.949	> .70

Note. Adapted from Renzulli, et al., 2009.

[a]In the combined scales, a mathematics item was removed, leaving a final math scale with 10 items; separate fit indices were not reported for the final 10-item model.

According to several researchers (Bentler, 1990; Dimitrov, 2012; Hu & Bentler, 1999), values for RMSEA < .08, χ^2/df < 5.0, CFI > .90 and TLI > .90 indicate good fit to the model, although Dimitrov indicates a χ^2/df < 2.0 as showing good model fit. However, decisions about overall model fit should be based on a joint assessment of all indices (Hu & Bentler) and not just one or two, with CFI, TLI and RMSEA being especially useful in practical applications (Dimitrov). Thus, the CFA indices for the SRBCSS-III are consistent with good model fit. The selection of the final items appears

to yield the most parsimonious model using CFA, which supports that the successful removal of redundant items or misfitting items, for example, items highly correlated with one another or with unacceptable factor loadings (Nazim & Ahmad, 2013), was accomplished. In the current study, these same data were evaluated using the Rasch rating scale model to answer Research Questions 1, 2 and 3.

Renzulli et al. (2009) hypothesized convergent validity between student ratings on the SRBCSS-III and the grades students earned in the content areas. Indeed, the authors report strong correlations between student grades and teacher ratings for reading, mathematics and science. Moreover, the authors present a concern with the strongest correlation between science and mathematics, which at greater than .85 could present a discriminant validity concern between the two scales. As well, for the technology scale, the ill-defined nature of what constitutes a technology grade made this area more difficult to interpret; and, almost one-third of the responses did not include a grade for technology. Renzulli et al. suggest that further research in discriminating between the mathematics and science domains and defining technology is needed for future editions of the scales.

Considerations in Measurement and Scale Development

Bond and Fox (2001) proposed that the key question in the analysis of data from instruments considers how well a theoretical intention has been empirically realized. In the context of this study, the theoretical intention of the instrument used is to identify students for gifted education services or placement into a developmental setting for growing their talent (Renzulli & Smith, 2010).

Wolfe and Smith (2007a) proposed a three-step process of defining and measuring attributes. First, observations of characteristics and behaviors of the intended targets are recorded (here, they are a number of the observable characteristics and behaviors from the theories in the previous section; e.g., see Table 5). Next, categories of observations are assigned descriptors that express how much or to what extent an individual possesses or exhibits the characteristics or behavior the researcher intends to measure. Finally, a measure is constructed from the condensed set of observations for a person.

The measure that is constructed positions a person at a location along a continuum of the attribute being measured (Bond & Fox, 2007; Osterlind, 2010). The measure represents a quantitative manifestation of the extent to which a person possesses some amount of the theoretical construct under consideration. From the magnitude of the measure, inferences and decisions are made about the persons that are measured. Because important decisions will be made based upon the measure, it is important that measures be reliable and allow for valid inferences regarding the persons being measured (Dimitrov, 2012; Messick, 1995b). The next sections contextualize these important concepts.

Aspects of validity. Validity refers to the interpretation of the data obtained from measurement and not to the instrument used in the measurement procedure (Messick, 1995b). Messick (1989, 1995a) and Kane (2001) maintained that validity is an evaluation argument–a judgment of the degree to which support for interpretation has been substantiated. The argument should be based on multiple sources of evidence, a

theoretical rationale and a clear understanding of relevant frameworks. In Messick's (1989) unified construct-based model of validity, six aspects of validity are discussed, of which four are particularly relevant to the current study: content validity, structural validity, substantive validity, and generalizability validity. That is not to say that additional aspects of validity are not important, but the current study will evaluate only the four aforementioned in the context of Rasch analysis.

Content validity. Content validity provides the evidence of content applicability and the representativeness of the construct being measured (Messick, 1989). In evaluating content validity, showing that an instrument's items are representative of the range of tasks, behaviors, abilities or other characteristics that are needed to perform in the area being measured would provide evidence for validity, for example (Dimitrov, 2012). Validity evidence can also be obtained through expert review of items or even from examinees completing the instrument. Additionally, content validity evidence is provided when a developer shows the extent to which the universe of potential items is reflected in an instrument's item; for example, by creating an instrument blueprint. Further support for the content aspect of validity is shown when experts' judgments are used to evaluate the readability and fairness of an instrument's items.

One component of the content aspect of validity surrounds the technical quality of the items on an instrument (Dimitrov, 2012). This can be evaluated through correlations between item-level/item-level and item-level/instrument-level scores (e.g., Pearson correlation and item-total correlation). In the context of the current study, the point-measure correlations were estimated using Rasch measurement, which is analogous to the

item-total correlation in classical test theory. Positive point-measure correlations provide evidence for content validity (Linacre, 2014a; Smith, 2003; Wolfe & Smith, 2007b).

Structural validity. Structural validity relates to the alignment of the scoring structure of an instrument to the structure of the construct around which the instrument is designed (Messick, 1995a). Specifically, the factor structure of the instrument as well as the appropriate selection of items within factors is evaluated to provide evidence for structural validity. In Rasch analysis, principal component analysis (PCA) can be used to provide such evidence (Linacre, 2014a), which involves the evaluation of correlations within the standardized residual variance (the unexplained part of the data). As Bond and Fox (2007) noted, it is really an examination of the extent to which the variance is explained by the measure itself.

Substantive validity. Broad in nature, one component of substantive validity that is particularly salient in Rasch-based arguments is scale functioning (Dimitrov, 2012). Substantive validity (Messick, 1989, 1995a, 1995b) refers to the observed consistency and patterns in responses of examinees or raters. Wolfe and Smith (2007b) related this to assessing the mean-square statistics, step calibrations and step difficulties, for example, in a Rasch analyses. Additional substantive validity evidence is provided by examination of item hierarchy through Rasch analysis and the item-separation index.

Generalizability validity. Generalizability validity evidence shows that the properties of the scoring and the interpretations of the scores reflect invariance across all groups to whom an assessment is administered (Dimitrov, 2012). Among others, one type

of evidence evaluated to support an argument for generalizability validity includes the absence of differential item functioning, which will be examined in the current study.

Reliability. Reliability is a necessary but not sufficient condition in assessing validity, it does signify additional evidence for validity. Reliability refers to the extent to which scores obtained from an instrument are free from random error (Dimitrov, 2012). In classical test theory, it is an expression of the correlation between observed score and true score, which suggests the notion that reliability expresses two ideas: error in measurement and replication for its estimation (Osterlind, 2010). In measurement, reliability increases as does the consistency of measurement and its increasing degree of accuracy.

In Rasch measurement, item reliability is a measure of the extent to which the items on an instrument can be precisely located along the latent variable. Low values suggest that the items are not representative of a wide range of difficulty or that the sample size was too small (Bond & Fox, 2007; Linacre, 2014a; Smith, 2003). Reliability is inextricably linked to the person-separation index, which will be evaluated in the current study.

Unidimensionality. Beyond the construct being measured, there is likely to be an additional element that is tapped in responding to items on an instrument (Dimitrov, 2012). The question of the effect of this dimension on the measure and the extent to which it influences the measure is an important aspect to review.

In constructing an instrument that can be evaluated using Rasch models there should be an underlying unidimensionality to the construct (Andrich, 1988; Bond & Fox,

2007; Dimitrov, 2012). That is not to say that there cannot be hints of additional dimensions, but the Rasch model evaluates data as if they are unidimensional. Fit statistics can be used to evaluate whether there is any conceptual multidimensionality that should be addressed with the instrument (Linacre, 2014a). Additionally, Rasch-PCA can be used to evaluate dimensionality. Fundamentally, Linacre suggested that the presence of multidimensionality may not be a concern in some circumstances; Bond and Fox (2001) suggested that "it is a matter of degree and not kind" (p. 103). If the data reflect misfit, review the purpose and use of the instrument, or carefully evaluate oppositional items in the PCA. At any rate, an evaluation of dimensionality yields additional evidence for the reliability and validity of an instrument.

Rating scales. Dimitrov (2012) suggested that a rating scale represents a set of ordered-category items each of which allows respondents the opportunity to indicate the level of their attitude, satisfaction or perception related to some construct. There are several types of rating scales, with Likert scales, Likert-type scales, frequency scales, and visual analog scales being some of the most widely-used. In the context of the current study, the literature on Likert-type scales and Likert-type items was reviewed.

Likert-type scales. A true Likert scale consists of Likert items, and the scale and items have the characteristics as described by Uebersax (2006).

1. The scale contains several items.
2. Response levels are arranged horizontally.
3. Response levels are anchored with consecutive integers.

4. Response levels are anchored with verbal labels that represent more-or-less evenly-spaced gradations.
5. Verbal labels are bivalent and symmetrical about a neutral middle.
6. The scale always measures attitude in terms of level of agreement or disagreement to a target statement.

The SRBCSS-III does not adhere to components three or five, and the options are stated in terms of indicating increasing levels of observation rather than levels of agreement with a target statement. However, according to Dimitrov (2012), there is not a need to strictly adhere to all criteria when representing items as Likert-type items as long as the fundamental interpretability remains true to the Likert scale methodology. Thus, the SRBCSS-III is better represented as a Likert-type scale comprised of Likert-type items.

Likert and Likert-type scales may be subject to several biases (Dimitrov, 2012), including (a) *central tendency*, whereby respondents avoid the extreme categories; (b) *acquiescence*, which occurs when respondents over-agree with the statements presented; and (c) *social desirability*, in which case respondents attempt to cast a generally favorable light in their rating. Acquiescence can often be mitigated with the use of negatively- and positively-worded items, but it is more difficult to deal with central tendency or social desirability.

Categories. Much debate continues around the number of response categories that should be used in constructing a Likert or Likert-type item. Pearse (2011) suggested that there are trade-offs in choosing to use many categories as there are in choosing to use few categories. With a greater number of categories there is a cognitive burden on

respondents to distinguish differences between the options, trivialization of the categories may occur, and establishing cognitive reference points may result in the use of only a few of the category options. On the other hand, a higher granularity might yield a greater meaningful variance, improve reliability and diminish the amount of missing data. With a lesser number of categories, however, the respondents may be able to finish more quickly, but this might come at the expense of bias in selecting responses or missing data where respondents cannot locate an item matching their level of agreement. In the current study, a category structure analysis was performed to evaluate the successful functioning of categories.

Rasch Measurement

The Rasch model is a probabilistic mathematical model that overcomes some of the measurement challenges classical test theory presents (Bond & Fox, 2001, 2007). Rasch models are mathematical models that construct quantitative measures from observations that are qualitative in nature (Bode & Wright, 1999) through the conversion of raw scores into linear measures (Iramaneerat, Smith, & Smith, 2008). In doing so, the models provide a probabilistic expectation of item and person performance when a single construct underlies a measure (Bond & Fox).

The family of Rasch models has characteristics that are shared between them, which in some cases is why many researchers separate the Rasch family from other measurement theory models such as item-response theory models (Osterlind, 2010). Before discussing some of the technical aspects of Rasch models, some of the characteristics of Rasch modeling that in large part underlie its advantages are presented.

Model characteristics. Rasch models are *stochastic*, which refers to the probabilistic expectation of randomness within the data (Bond & Fox, 2007). Where data are too predictable–areas within the data that show too little randomness; i.e., a Guttman pattern–the performances of respondents appear to be more different than they actually are. In the opposite case–the data are excessively random–the performances of the respondents appear to be more similar than they actually are by collapsing the measurement system. It is the latter that is more disruptive to the measurement (Linacre, 2002).

Bond and Fox (2007) described the importance of *model fit*. Rasch models are based on tests of fit, in which the data are tested for fit against the model rather than finding a model that fits the data. This is important to the extent that the advantages of fundamental measurement afforded by the Rasch model (i.e., equal interval measures resulting in conjoint measurement) exist because data must conform to the model. Where data do not fit, it is essential for the researcher to investigate the instrument, responses or persons for more information.

Rasch models provide measures of the amount of a latent trait of persons and items that are invariant, which results in *specific objectivity* (Iramaneerat et al., 2008). In other words, the estimated difference in ability between two persons is independent of the difficulty of any certain items used to compare them, and the difficulty of the items is independent of the ability of the persons responding to them (Dimitrov, 2012; Irwin, 2007).

The concept of *invariance* is related to the notion of specific objectivity. Fundamentally, the order of persons according to their level on the trait and the order of the items according to their difficulties is invariant: a person with more of a trait should always have a higher probability of correctly endorsing (or endorsing at a higher level) than a person with less of the trait (Andrich, 1988). Moreover, more difficult items should always have a lower chance of being endorsed correctly than easier items no matter the who attempts the items (Iramaneerat et al., 2008; Linacre, 2014a)

Each measure has its own *standard error* associated with its estimate. This is in contrast to classical test theory, where a single standard error is provided for the range of scores. The advantage is the precision of estimates of ability along most of a continuum even while the extreme scores have greater standard error. A second advantage of multiple standard errors is that it gives developers the opportunity to readily see the positions along the measure continuum where test information is at its lowest and fill in the area with new items. Finally, estimates of reliability are more accurate owing to the range of standard errors associated with ability estimates using Rasch models (Iramaneerat et al., 2008; Wolfe & Smith, 2007b).

Lastly, the Rasch model scales persons and items onto the same scale, and the scale is *interval* in nature. This allows for additive conjoint measurement, which is not available from raw scores alone (Perline, Wright, & Wainer, 1979). Measures from the Rasch conversion onto a common *logit* (log odds unit) scale can be directly compared between two persons, two items or between persons and items.

Model framework. The Rasch model is a probabilistic model developed by Danish mathematician Georg Rasch (1960) and extended by Wright, Andrich, Samejima, Masters and others over the last fifty years into a wide-reaching and extensive family of probabilistic models for measurement (Ostini & Nering, 2006). On the surface, the model appears theoretically simple–and, indeed it is (Osterlind, 2010). As discussed in the preceding sections, a developer identifies items aligned to a construct, administers an instrument containing the items and then evaluates the fit of the data to the Rasch model. However, as Osterlind pointed out, the practice is more challenging than the theory:

- stimuli must be focused on a single construct,
- examinees must employ only the anticipated cognitive processes,
- maximal effort must be exerted by examinees, and
- unidimensionality and local independence are necessary.

Nonetheless, if the assumptions hold and data fit the model, Bond and Fox suggested that the model is a "compact, efficient, and effective form of what measurement in the human sciences should always be like" (Bond & Fox, 2001, p. xv).

Rasch rating scale model. The Rasch rating scale model (RSM) is used in the case of polytomous items that have the same category structure across all items (Andrich, 1988). The RSM can be expressed mathematically as (Dimitrov, 2012)

$$\ln\left(\frac{P_{nik}}{P_{ni(k-1)}}\right) = \theta_n - (\delta_i + \tau_k)$$

where

ln(.) represents the natural logarithm,

P_{nik} is the probability that person n would respond in category k when answering item i,

$P_{ni(k-1)}$ is the probability that person n would respond in category $k - 1$ when answering item i,

θ_n is the trait score of person n on the logit scale,

δ_i is the difficulty of item i, and

τ_k is a threshold indicating the impediment to being observed in category k relative to category $k - 1$.

From the equation above, the item difficulty δ_i represents the location on the scale at which a respondent has a .50 probability of responding in either of the two extreme categories. From the equation, category probability curves can be constructed as shown in Figure 3, which reflect the probabilities of responding within a category k given trait level θ.

Figure 3 reflects that the easiest category to endorse for those with low ability is category one, while the most difficult to endorse is the category five. The opposite is true for those with high ability on the right side of Figure 3. What the RSM probability curves in Figure 3 show is that as the amount of latent trait increases the probability of selecting a higher-level category increases–this is an essential component of a functioning category structure (Smith, 2003). Moreover, the curves reflect that the fundamental measurement in the Rasch model is rooted in the item difficulty and the person ability. Figure 3 shows, for example, that a person with an ability level that is one logit below the difficulty of an item, or $\theta = -1$, then the probability of endorsing the item in category four would be just

under .1, while the probability of endorsing the item in category two would be almost .4. Indeed, this represents the foundation of the Rasch model. More on the use of the category probability curves, their evaluation and diagnoses of the SRBCSS-III utilizing the curves and their thresholds will be presented in the Discussion and Conclusions section later.

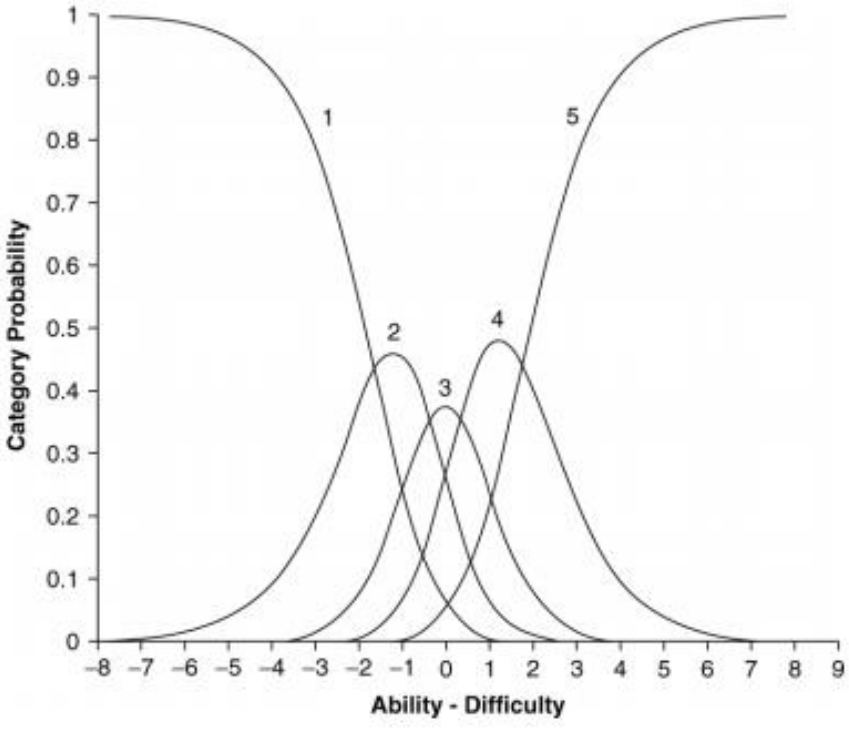

Figure 3. Category probability curves. Figure represents a polytomous item on a five-point scale presented as measure relative to item difficulty.

CHAPTER THREE: METHODS

Review and Permissions

This study uses existing data sets that were collected as part of the operational activities of identifying students in Grade 3 and Grade 4 for access to enrichment activities delivered in a gifted and talented setting (the local data sets) and data provided by the authors of the SRBCSS-III (the SRBCSS data sets). The George Mason University Office of Research Integrity & Assurance and Institutional Review Board granted exempt status for review owing to the existing nature of the data sets. Permission was separately granted by the local school district from which the data were obtained to use the local data sets with two stipulations: a) that no personally identifiable information was made available by the school departments involved or the independent vendor on whose platform the data were collected, and b) that a full electronic copy of the final research study be provided to the district after successful defense at George Mason University.

Instrument

The Scales for Rating the Behavioral Characteristics of Superior Students (Renzulli et al., 2013) were originally published in 1978 to assess the characteristics of high-ability students for whom gifted programs were appropriate. The updated edition of the revised SRBCSS-R, the SRBCSS-III, was used in the Rasch analyses conducted in this study. Originally developed in 1971, the SRBCSS has undergone two revisions, the

first in 2002 and the second in 2010 when four additional scales were added to the 2002 revised scales. Thus, the current edition of the SRBCSS-III contains fourteen scales–the ten that were on the revised edition and the four domain scales added in 2010. Although the four most commonly-used scales in identification are the leadership, creativity, motivation and learning characteristics scales (Renzulli & Smith, 2010), this study evaluates data collected in the 2010 validation study of the domain scales reading, mathematics, science, technology scales and the data collected in an operational administration using the reading, mathematics, learning characteristics and motivation scales.

All items on the SRBCSS-III use six category options, an example of which is shown in Figure 4; full sample copies of the six operational scales used in the study are shown in Appendix A. The resulting scales are comprised of essentially Likert-type items that in general methodology conform to the Likert scale concept even though three generally accepted characteristics of strict Likert items are absent according to Uebersax's 2006 summary of the components of strict Likert items.

1. The scale contains several items.
2. Response levels are arranged horizontally.
3. Response levels are anchored with consecutive integers.
4. Response levels are anchored with verbal labels that represent more-or-less evenly-spaced gradations.
5. Verbal labels are bivalent and symmetrical about a neutral middle.

6. The scale always measures attitude in terms of level of agreement or disagreement to a target statement.

Specifically, the SRBCSS-III does not adhere to components three or five, and the options are stated in terms of indicating increasing levels of observation rather than levels of agreement with a target statement. However, according to Dimitrov (2012), there is not a need to strictly adhere to all criteria when representing items as Likert-type items as long as the fundamental interpretability remains true to the Likert scale methodology.

Figure 4. Example Likert-type item with descriptors from SRBCSS-III.

Teacher instructions for the scales include a statement indicating that the scales are designed to obtain teacher estimates of a student's characteristics in the area of each scale, and that each item on a scale should be considered separately from other items on the scale. Teachers are instructed that a rating "should reflect the degree to which you have observed the presence or absence of each characteristic" (Renzulli et al., 2013, p. 1). For each item on the scales, teachers indicate the extent to which they have observed the characteristic described in the item stem. As an example, as shown in Figure 4, a teacher will respond to the stem *The student eagerly engages in reading related activities* by

indicating his or her observation frequency of this characteristic as *never, very rarely, rarely, occasionally, frequently,* or *always*.

A challenging aspect of the SRBCSS-III is that the response categories imply that opportunities for observation of a characteristic must have been made available in the classroom setting in order to endorse an item at all. For example, on the mathematics scale, item three asks teachers to respond to whether the student *enjoys challenging math puzzles, games and logic problems*. In order to endorse this item using the categories shown in Figure 4, there is an assumption that the student must have been provided an opportunity to engage in such an activity. Specific instructions for this apparent problem were provided to teachers and are discussed later.

The motivation and learning characteristics scales each contain eleven items, while the mathematics and reading scales contain ten and six, respectively. The estimated burden for the full instrument has been estimated at 40 minutes (Jarosewich et al., 2002), but no recommendations for individual scales are provided. In the district in which data were collected, teachers in Grade 3 and Grade 4 have an approximate five week window in late spring to complete the rating scales for their students.

Scoring on the SRBCSS-III is accomplished by assigning a point value of one to six across the rating categories, where *never* = 1, *very rarely* = 2, *rarely* = 3, *occasionally* = 4, *frequently* = 5 and *always* = 6 for each item. Sum totals for each category are then added across all categories to obtain a total score for a domain[1] (e.g.,

[1] Throughout, "domain" is used when referring to the construct of a content area or latent trait, while "scale/s" is used when referring to the rating scale instrument for a domain.

reading or mathematics). Explicit directions are provided in the SRBCSS-III manual to avoid summing scores across domains in an attempt to obtain a grand total score for students (Renzulli & Smith, 2010), as students might be identified in a single domain and not in others, which is consistent with theories of giftedness and identified widely as best practice (Robinson et al., 2007; VanTassel-Baska, 2008). Lohman (2005) and Renzulli and Smith recommend that local districts calculate local norms for each domain and use these local norms and student domain scores in the identification process, which is well-described in the most recent administration manual. However, the district in which this study was performed, indeed, summed scores across pairs of domains (reading and learning characteristics; mathematics and motivation) to obtain two grand total scores for identification. Because the scales are intended to identify students in separate domains according to the talent development frameworks around which the SRBCSS-III were developed, it is counter to the use of the scales to sum scores across scales (Renzulli & Smith).

Participants and Setting

Two sets of data were used in this study. The first set contains data that were collected in a local school district (the local data), which was used to answer Research Questions 2 and 4 in the current study relating to analyses of the SRBCSS-III in operational administrations. The second set (the SRBCSS data) contains data provided by the SRBCSS-III authors, which contains the data that were used for the validation study of the new scales added to the SRBCSS-III (Renzulli et al., 2013). These data were used to address Research Questions 1, 2 and 3, which relate to the reliability, factor structure,

rating scale structure and validity of the scales added to the SRBCSS-III. The data sets and the participant characteristics for each set are separately described in the following sections.

Participants and data: Local data sets. Data for the local sample were provided by a mid-sized, suburban, county public school district in a Mid-Atlantic state from teacher ratings performed in the spring of 2013 and the spring 2014. The district has almost 16,000 students, with approximately 1,100 students in each Grade 3 and Grade 4 across 12 elementary schools, which are the grades for which ratings by teachers were completed. In the first year, 46 teachers rated the 2013 Grade 3 cohort, and these students were again rated in 2014 as Grade 4 students by 41 teachers. The 2014 Grade 3 cohort was rated by 48 teachers, 41 of whom were also teachers rating the 2013 Grade 3 cohort. Thus, a total of 94 teachers were involved in data collection over the two school-year period for which data were provided. The motivation and learning characteristics scales have been used in the district each year since they were adopted for first use in 2004, and the reading and mathematics scales have been used since they were added to the SRBCSS-III in 2010.

Local participants. The local sample consisted of students in Grades 3 and Grade 4, which are the two lowest grades for which the SRBCSS-III is recommended (Renzulli & Smith, 2010). The number of students in the 2013 Grade 3 cohort that were also rated in 2014 as Grade 4 students was 1,024 (93.1% of the 2013 Grade 3 cohort continued into 2014); two years of rating data are available for this sample of students. The 2014 Grade 3 cohort contained 1,097 rated students; one year of data are available for this sample of

students. Cohort characteristics for the students on whom ratings were collected are shown in Table 7. The cohort characteristics illustrate the nature of narrow demographic representativeness in the sample, which is discussed later as a limitation of the current study.

Table 7

Student Cohort Characteristics, Local Data Set, Percentages

Characteristic	2013 Grade 3 Cohort	2014 Grade 3 Cohort
Sex		
Male	51.0	48.2
Female	49.0	51.8
Ethnicity		
American Indian	<10	<10
Asian/Pacific Islander	<10	<10
Black	11.2	11.9
Hispanic	<10	10.0
Multi-Racial	<10	<10
White	76.5	73.7
Students with disabilities	11.8	<10
Economically disadvantaged	24.9	26.5
Limited English proficient	<10	<10
Identified for enrichment[a]		
End of Grade 3	33.5	
End of Grade 4	36.4	

Note. Data are suppressed for demographic groups representing less than 10% of sample.
[a]Students may be identified in one or more areas for targeted enrichment opportunities.

A census review of students is completed each year on the students in the rated grades to comply with the district's state-mandated master plan. This results in students identified for Grade 4 enrichment at the end of Grade 3 being re-evaluated for Grade 5 enrichment at the end of their Grade 4 year. Grade 3 is the first year of identification in the district, and a school-wide enrichment model for gifted and talented education services is used at the lower grades where formal identification does not occur. The local data used in this study were not collected specifically for the purpose of this study; the rating of students by teachers using the SRBCSS-III is a component of the standard identification process for providing specialized enrichment opportunities in Grade 4 (for Grade 3 students) or Grade 5 (for Grade 4 students). Students with profound intellectual disabilities that participate in the district's functional skills education program were not rated, but students that were otherwise identified for special education services alongside the regular education program were included in the census review.

Teachers worked independently to rate students they taught in mathematics or reading. Teachers of mathematics completed the motivation and mathematics scales for their students, while teachers of reading completed the reading and learning characteristics scales for their students. No teachers teach both mathematics and reading, so the maximum number of SRBCSS-III domain scales they completed was two, and the mean number of students rated in 2014 by a teacher was 50 for Grade 4 teachers and 46 for Grade 3. Grade 3 teachers in 2013 rated an average of 47 students. The maximum raw score that could be attained was 126 on the mathematics-motivation combination, while the maximum raw score that could be attained was 102 on the reading-learning

characteristics combination. Because the minimum category rating is assigned a value of one, the minimum score on either is the sum of the questions, or 21 for mathematics-motivation and 17 for reading-learning characteristics.

Teachers were not provided formal training on the rating scales, but the central office of the district did require participation in a discussion session at each school relating to the purpose of completing the scales and some things to think about when considering their students. To facilitate this, central office learning specialists for gifted and talented education visited each elementary school to discuss with the teachers that the ratings should consider classroom observations of the behaviors on the scales and not things, for example, such as whether students completed homework, had large numbers of absences or were already in an accelerated classroom. Although this was informal, the consensus was that the short discussions provided insight into setting parameters for what teachers should be thinking about as they completed the scales. Teachers likely attained at least a moderate proficiency at understanding how the scales should be used, although only a few with sufficient experience likely understood the use of the scales at an advanced or expert level. The rating scale training exercises provided in the *SRBCSS-III Technical and Administration Manual* (Renzulli & Smith, 2010) were not completed system-wide, although small self-directed groups of teachers may have discussed or completed the exercises, as they were provided to the lead teacher in each grade with the materials for raters.

Local data. Creswell (2012) discusses the value of nonprobability sampling in selecting a sample for research that represents a characteristic or set of characteristics that are of interest in answering research questions. Here, the question surrounds various aspects of validity, reliability, and other features of a rating scale for identification in gifted and talented education for elementary students. Thus, a convenience sample of Grade 3 and Grade 4 students was selected for participation in this study.

The sample is sufficiently large to address the requirements of Rasch analysis sample sizes to ensure the stability of measures (Linacre, 1997). Considering a well-targeted sample, the recommendation for minimum sample size to obtain item calibration stability within ±1 logit, 99% CI is 27 persons, while obtaining stable item calibrations with ±0.5 logit, 99% CI is 108 persons. Moreover, in light of the targeted nature of the sample of Grade 3 and Grade 4 students for whom the SRBCSS-III was designed, the sample of greater than 1,000 persons exceeds even the confidence demanded of high stakes testing circumstances according to Linacre (2007) and Kruyen, Emons, and Sijtsma (2012).

Teachers independently rated students over a five week time period leading up to and including the last week of the 2013 and 2014 school years on all of their Grade 3 and Grade 4 students. Teachers were free to complete ratings over several sittings, and where teachers had a student for whom an item was particularly difficult to endorse they were told to seek input from another teacher with familiarity with the student.

Owing to the category responses available for endorsement, all of which require that an observation of a student engaged in a situation where the behavioral characteristic could have been made, teachers were instructed to leave an item blank if a student had not engaged in such a way that there had been an opportunity to observe the characteristic. Recalling the example cited earlier, if a student had not been provided opportunities to engage in challenging math puzzles, math games or logic problems, it would be misleading for a teacher to endorse any of the categories. This was not considered a concern, as the Rasch rating scale model utilizes joint maximum likelihood estimation, which Linacre (2004) describes as flexible with regards to missing data.

Teachers completed the student ratings within an online scoring environment platform provided by a district vendor. To complete the rating, a teacher selected a single student from a prepopulated roster and selected a rating scale, which was presented with an onscreen layout similar to that of the paper format. That is, the items were listed vertically and the category options were presented horizontally. Responses were selected by clicking within checkboxes aligned with the items. A subsequent check placed in a box for an item caused any earlier selection option to be unchecked automatically. Upon selecting to submit a student's ratings, the online platform prompted to warn of any missing items and allowed teachers to confirm submission or return to the rating environment. Teachers could return to change students' ratings through the five week period during which the rating scales were open.

Data collection. Student data were extracted by the district's online platform vendor and provided as Excel spreadsheet files. The files contained student IDs, rating scale test identifiers, and scale responses for each cohort of students. A separate file contained demographic information for students, including sex, ethnicity, disability status, economically disadvantaged status, limited English proficiency status, and student ID for all three groups of students.

Data cleansing involved removing 2014 Grade 4 students for whom 2013 Grade 3 ratings were not available and removing 2013 Grade 3 students for whom 2014 Grade 4 ratings were not available. As well, students were connected to their scale scores and demographic information through their student ID number, as no student names were provided in the data files.

Scale responses in the student data sets were provided as numerical values, where 1 = *never*, 2 = *very rarely*, 3= *rarely*, 4 = *occasionally*, 5 = *frequently*, and 6 = *always*. The data files were recoded to assign zero to the lowest observation level (i.e., 0 = *never*, 1 = *very rarely*, 2 = *rarely*, 3 = *occasionally*, 4 = *frequently*, and 5 = *always*) once they were merged for use with Rasch analysis software, which Bond and Fox (2007) offer as a recommended practice for analysis with common Rasch software. Missing data were presented as empty cells.

Participants and data: SRBCSS data sets. The participants in the reliability and factor structure studies related to the development of the SRBCSS-III can be separated into two groups: a) experts in the field of gifted and talented education, teachers and professors in the domains to be included, and resource specialists in the areas of

mathematics and reading; and b) the teachers and students that administered the field test of the proposed new scales in mathematics, reading, science and technology (Renzulli & Smith, 2010; Renzulli et al., 2009). The experts were involved in identifying characteristics and corresponding items to be included on the various scales to be field tested, and the teacher and student samples were involved in the later factor structure, validity and reliability studies of the constructed scales. The data provided by the authors of the SRBCSS-III for use in this study were the data collected in the field tests by teachers and students.

SRBCSS Participants. Renzulli et al. (2009) reported that 187 teachers rated 726 students from 140 schools in districts self-described as urban (26%), suburban (64%) or rural (10%) participated in the field testing of the four new scales incorporated into the SRBCSS-III. One hundred twenty-two schools offered gifted programs. Teachers to whom the field test scales were mailed were asked to complete the scales on every fifth student on their roster. Additional demographic characteristics of this sample of students is provided in Table 8. Similar to the local data sets, a narrow demographic representativeness can be seen with the SRBCSS sample. The Renzulli et al. report does not indicate if responses were received from all 140 schools to which the initial mailing was made.

Table 8

Student Characteristics, SRBCSS-III Authors' Sample

Characteristic	Percentage
Sex	
Male	52
Female	48
Ethnicity	
Native American	1
African American	8
Hispanic	7
White	80
Enrolled in gifted programs	31

SRBCSS data. The reliability and factor structure study data were emailed to me by Del Siegle, the author assigned as the custodian of the records for the SRBCSS-III domain scales study according to Renzulli (personal communication, June 4, 2014). Data were received in a single flatfile spreadsheet file.

In the data file, data for 73 potential rating scale items were recorded from the field test, although the final domain scales used only 30 items total (6 Reading, 10 mathematics, 7 science and 7 technology). For this study, only the data for the 30 items selected for the operational subscales were used; the list of items is provided in Appendix C.

Rasch Analysis Procedures

Wright (1977) suggested that the Rasch model is the manifest example of the assumption that the unweighted sum of right answers by a person and the unweighted sum of correct endorsements to items is all that is needed to measure a person and

calibrate items, respectively. If the data fit the Rasch model, Wright asserted that the placement of a person on a measurement scale is entirely a function of observable data. The sections that follow describe the model fit indices and evaluative examinations of instrument data and analysis output that were made to determine the extent to which the data in the SRBCSS-III data set fit the Rasch rating scale model and provide evidence to answer Research Questions 1, 2, and 3 as well as how the local data were evaluated to answer Research Questions 2 and 3.

Rating scale model. The Rasch rating scale model, RSM, was used in the analyses of both the SRBCSS data set and local data sets. The computer program *WINSTEPS* (Linacre, 2014b) was used to analyze the data in both data sets. The SRBCSS-III uses the same rating scale across all items. The RSM is one of the most widely-used Rasch models for polytomous data, and it is the recommended model when all items share a common rating scale (Linacre, 2000). In fact, Linacre indicated that strong evidence would be needed to use a model other than the RSM (e.g., partial credit model) where all items have the same rating scale. Importantly, the RSM has the benefit of being robust to missing and accidental data in addition to dealing with situations where a few items have underutilized categories relative to other items. The *WINSTEPS* program has a default setting to utilize the partial credit model, which was manually changed to the RSM in running the program. However, Linacre (2014b) noted that when all items share a common rating scale, the partial credit model is, in fact, the RSM.

RSM analysis: Validity and reliability evidence. Smith (2003), Dimitrov and Smith (2006), and Bond and Fox (2007) extensively characterized the evaluations and diagnostics to employ in evaluating the fit of data to the RSM and building evidence for reliability and validity. Linacre (2002) and Bond and Fox further described the evaluations of rating scale structure and category effectiveness using the RSM. These analyses will be used to answer Research Questions 1 and 2.

Person separation and reliability. Person separation indices were evaluated using generally accepted parameters for such measures provided in the literature (Linacre, 1997, 2014a), although the essential guideline for cut-off parameters is whether or not the instrument distinguishes a sample into enough levels for a particular purpose. Person separation indicates the extent to which an instrument's scale discriminates well between persons (Smith, 2003). The real person separation was used for this study, as it accounts for any error that arises from model misfit (Bond & Fox, 2007; Smith), and it is calculated as the ratio of the square root of the variance explained by the model to that of measurement error. Real person separation greater than 2.0 with person reliability greater than .80 implies that the instrument is likely sensitive enough to accurately classify between those of high ability and those of low ability on the instrument. Person reliability was evaluated in conjunction with person separation.

The person reliability expresses the probability that persons on the high range of ability do, indeed, have a high ability, while those of low ability are, indeed, likely to be found on the lower range of ability on another measure of the same variable. Linacre (2014a) suggested that person reliability is generally dependent on a wide range of

abilities for participants, the number of categories per item and the targeting of the persons.

Item separation and reliability. Item separation confirms the item difficulty hierarchy. Item separation is particularly useful in providing evidence for content validity (Linacre, 2014a). Values of item separation lower than 3.0 with item reliability less than .90 implies a lack of items at a wide enough range of difficulties to provide evidence for content validity; i.e., the items potentially do not offer a range of difficulties that cover the range of the construct and might not represent a well-defined variable (Smith, 2003).

Item reliability is a measure of the extent to which the items on an instrument can be precisely located along the latent variable. Low values suggest that the items are not representative of a wide range of difficulty or that the sample size was too small (Bond & Fox, 2007; Linacre, 2014a; Smith, 2003).

Item statistics. Linacre (2014a) and Smith (2003) suggested that the point-measure correlations should be reviewed prior to evaluating any misfit using Outfit and Infit statistics. Point-measure correlations represent the correlation between person measures and their responses to an item, and point-measure correlations are robust to missing data. Point-measure correlations can be valuable in providing support for convergent and divergent validity, and were evaluated as additional evidence for content validity.

Bond and Fox (2007) asserted clearly that the task of the Rasch model is not to account for the data but is instead a model to describe how fundamental measurement should appear. Because models do not hold in practice, misfit can occur in the

measurement model at both item and person level when the data do not perfectly fit the Rasch model (Smith, 2003). Moreover, the probabilistic nature of the Rasch model does not expect perfect fit, and too little variation–as in a Guttman response pattern–would, indeed, reflect misfit in the Rasch model according to Bond and Fox.

In the Rasch model misfit can be diagnosed using the Infit mean-square (MnSq) and Outfit MnSq statistics. These statistics are chi-square, χ^2, ratios rooted in the squared standardized residuals (Dimitrov, 2012; Wright & Masters, 1982). The standardized residual is calculated as $Z = (X - E)/\sqrt{\mathrm{VAR}(X)}$, where X is the observed score, E is the Rasch model expected value and VAR(X) is the variance of the observed scores. The sum of the squared standardized residuals is a χ^2 statistic. When data fit the model, Infit MnSq and Outfit MnSq have expected values of one.

The Outfit MnSq statistic is sensitive to outliers, and Dimitrov (2012) suggested that it is sensitive to unexpected rather than misfitting responses. Values of Outfit MnSq less than 1.0, or overfit, indicate the model predicts the data too well and inflate summary statistics such as reliability. Values of Outfit MnSq greater than 1.0, or underfit, reflect noise in the data, which can degrade measurement. Underfit is a more immediate threat to measurement.

The Infit MnSq statistic signals unexpected response patterns for in-target measures; that is, a person is responding in a more haphazard way than expected (Bond & Fox, 2007). As with Outfit MnSq statistics, values greater than 1.0 reflect underfit while values less than 1.0 represent reflect overfit. In the case of Infit MnSq, underfit is the larger threat to validity, but it is more difficult to diagnose (Linacre, 2014a).

Table 9

Acceptable Fit Statistics and Their Meanings

Mean-Square	Interpretation	Variation
MnSq > 2.0	Degrades or distorts measurement; haphazard response patterns	Greater than expected
1.5 < MnSq ≤ 2.0	Unproductive but not degrading; noticeably unpredictable patterns	Greater than expected
0.5 ≤ MnSq ≤ 1.5	Productive for measurement	Stochastically expected
MnSq < 0.5	Less productive for measurement but not degrading; may inflate reliability; too predictable–may be influenced by a constraining, restricting dimension	Less than expected

Note. Adapted from Linacre, 2014a; Dimitrov, 2012.

More generally, the Infit MnSq and Outfit MnSq are transformed into standardized MnSq *t*-statistic, referred to as the standardized *z*-statistic, or Zstd (Bond & Fox, 2007; Dimitrov, 2012). This statistic, which follows a normal distribution, is often used for diagnosis of fit. However, Linacre (2014a) indicated that the Zstd statistic only needs referenced if MnSq values are unacceptable. Recommended values of MnSq and interpretive information as used in this study are shown in Table 9.

General keyforms. The general keyforms were evaluated for the SRBCSS data to determine whether meaningful constructs were apparent in the item hierarchies on the scales evaluated to answer Research Question 2. An item hierarchy in accord with conceptualizations of a construct provide evidence for content validity.

Item-person map. Bond and Fox (2001) described the analysis of data using the Rasch model as "an estimate of what our construct might look like if we were to create a ruler to measure it" (p. 8). Fundamental to visualizing that the measurement ruler is able to measure along the continuum of abilities and item difficulties is the item-person map. In the case of Rasch measurement, the ruler is constructed of equal interval measures called logits, along which items and persons can be located on the same scale (Andrich, 1988). The advantage is that the difference (e.g., in ability or difficulty) between any two equally-spaced locations (i.e., 3 - 2 = 5 - 4) on the logit scale is the same, and the measures are additive. For example, someone with a logit ability estimate of 3 has a 2.72 times greater odds of responding correctly to an item than someone with a logit ability estimate of 2 (3 - 2 = 1; $e^1 = 2.72$); this is also the increase in odds for a correct response in the case of a pair of examinees at 5 and 4 for the person with ability estimate 5. This information cannot be obtained in classical test theory.

Validity evidence is provided when the items on the item-person map spread along the continuum of the logit scale and that persons spread along the continuum of the ability estimates the scale also represents. Fundamentally, this examination was to determine whether the instrument provides "well-spaced items that [cover] a substantial length of the construct" (Green & Frantom, 2002, p. 27), which provides evidence for content validity.

RSM analysis: Category and rating scale functioning. Rating scale functioning was reviewed for the SRBCSS data to answer Research Question 3. Smith (2003), Linacre (2002; 2014a) and Bond and Fox (2001, 2007) provided well-grounded

diagnostics for rating scale effectiveness. Following an evaluation of rating scale use summary information, Linacre, Bond and Fox, and Smith recommended the evaluation of additional output for analyses, which are described below.

Average measures and coherence. To ensure a meaningful interpretation that higher measures imply higher ability on the SRBCSS data and the local data, an evaluation of the average measures across all categories was made. Average measures should advance monotonically along the rating scale. Coherence between measures and category observations was also evaluated. Coherence expresses the number of measures that were expected to produce observations in a category as a proportion of those that actually did. Additionally, coherence expresses the proportion of observations in a category that were produced by measures corresponding to the category. A general coherence threshold is that it is acceptable above 40% (Dimitrov, 2012; Linacre, 2002).

Outfit mean-square of categories. Outfit MnSq of categories was evaluated next to determine if the category was used in an idiosyncratic way or in some unexpected context; values greater than 1.5 or higher indicate a large amount of unexplained noise in the data and significant misinformation in the category.

Step calibrations. Andrich (1996) emphasized that as the measures of persons increase, the probability of observing a person in a higher category should increase as well. For example, for a person with low ability, the probability of observing the person in category zero must be higher than the probability of observing the person in category five. Visual examination of the probability characteristics curves was done to verify the

ordering of step calibrations–each category should appear to have a peak at some point where it is the most probable category to be endorsed.

Step difficulties. The number of response categories should be large enough to identify along a wide range of the variable on an instrument but small enough so that respondents can conceptualize substantive differences of meaning between the category labels (Linacre, 2002). The step difficulties between two categories k and $k - 1$ should advance by at least 1.0 logit, which indicates a meaningful dichotomy between label k and label $k - 1$. Step difficulties should advance by no more than 5.0 logits, or a loss of precision and information results.

RSM analysis: Rasch principal component analysis. Bond and Fox (2007) characterized that the existence of a unidimensional construct can be presumed when the largest amount of the variance is explained by the measure. Rasch principal component analysis of residuals (PCA) detects correlations within the standardized residual variance, or the unexplained part of the data (Linacre, 2014a). The purpose of Rasch PCA–to identify and explain variance after the contribution of the measure has been removed–is in contrast to the variable construction purpose of common factor analysis.

In Rasch PCA, potential groupings of items that might correlate strongly enough to be a secondary dimension are detected. In the current study, Rasch PCA was conducted to determine whether a) any groups of items formed a substantive secondary dimension, and b) the content of any item was such that there was evidence for deleting it from the final rating scale.

Differential item functioning. Evaluation of differential item functioning (DIF) is important to ensuring that valid inferences are made from an instrument. DIF indicates that one group of respondents is scoring in a different way compared to one or more other groups after adjusting for all respondents' overall abilities. This yields evidence for a generalizability validity argument.

In the current study, DIF was detected using the method of DIF contrast. In the method of DIF contrast, all persons are anchored at their measure estimates from the main analysis, while the item difficulties are unanchored. Then, item difficulties are estimated for each group. Finally, the difference between the item difficulties by groups is computed, which yields the DIF contrast. An acceptable interpretation of the absolute values of DIF contrasts places items in a *negligible* DIF category ($|\text{DIF contrast}| < 0.43$), a *slight-moderate* DIF category ($0.43 \leq |\text{DIF contrast}| < 0.64$), or a *moderate-large* DIF category ($|\text{DIF contrast}| \geq 0.64$) (Linacre, 2014a).

In the current study, male students, Caucasian students and non-economically disadvantaged students were used as reference groups in the DIF analyses with respect to sex, race and ethnicity, and economic disadvantage (ED). For race and ethnicity DIF analysis, students reporting as African American, Hispanic, Asian, Pacific Islander and Native American were placed in the focal group, while students receiving free or reduced price lunch benefits were placed in the ED focal group.

Limitations

Limitations in the current study include the training of the raters, which did not occur in a coordinated way across the local data set participants. It is plausible that some teachers had a low proficiency in understanding the process, while more experienced teachers likely had a much greater proficiency.

In addition, the demographic composite of the students or the teachers in the local data set was not representative of the larger population. The local data were collected in one suburban, Mid-Atlantic county with a Grade 3 and Grade 4 population of just about 2,400 students. Additional analyses with a more representative sample would be advantageous to lending support to the conclusions.

CHAPTER FOUR: RESULTS

The purpose of this study is to use Rasch analysis to evaluate the validity, characteristics of reliability, dimensionality, item selection, category structure and differential item functioning of several domain subscales on the SRBCSS-III (Renzulli et al., 2013), which are commonly used in identifying students for placement in gifted programs. This chapter presents the results of the Rasch analyses conducted to answer:

Research Question 1: Does Rasch analysis confirm the dimensionality and evidence of well-functioning retained items on the domain scales (reading, mathematics, science and technology) added to the SRBCSS-III?

Research Question 2: Does Rasch analysis provide evidence for reliability and validity for the domain scales (reading, mathematics, science, and technology) and the learning characteristics and motivation scales on the SRBCSS-III?

Research Question 3: Does Rasch analysis show optimal category structure for the SRBCSS-III?

Research Question 4: Is there evidence of differential item functioning for subgroups of students on selected scales of the SRBCSS-III?

The first analyses in the current study were completed using the SRBCSS-III authors' validation field study data for the retained items. Next, analyses on both the SRBCSS-III data (using only the retained items from the authors' validation study field test) and local Grade 3 data were conducted to address Research Question 2. The third set of analyses investigated the category structure of the SRBCSS-III using only the data for the retained items on the domain scales included on the SRBCSS-III. Finally, the last analyses used Grade 4 local data to answer Research Question 4. A summary of how the data were used in the analyses is shown in Table 10. Misfitting (Outfit MnSq > 1.5) and extreme score persons (maximum and minimum scores) were removed in all analyses.

Table 10

Summary of Data Used in Analyses in the Current Study

Research Question	Data Set Used	Analyses
1	Retained items from the SRBCSS authors' validation study field test data (math, reading, science and technology)	Dimensionality; item retention
2	Retained items from the SRBCSS authors' validation study field test data (math, reading, science and technology); local data (Grade 3 learning characteristics and motivation)	Validity, reliability of the SRBCSS-III retained items
3	Retained items from the SRBCSS authors' validation study field test data (math, reading, science and technology)	Category structure of SRBCSS-III retained items
4	Local data (Grade 4 mathematics, reading, learning characteristics, and motivation)	DIF in operational form of SRBCSS-III

Results of Analyses for Research Question 1

In the first set of analyses, the objective was to investigate the dimensionality of the four domain scales of the SRBCSS-III and to explore the retention of items from their field test item sets. Rasch principal component analysis (RPCA) was conducted on each of the field test item sets, which provided data from which the dimensionality aspect of Research Question 1 was addressed. In the second set of analyses, item fit was evaluated to investigate the retention of items for the operational scales devised from the initial set of 73 items on the field test. The results of these analyses are presented in the next two sections.

Dimensionality analyses. Each of the four domain scales was separately evaluated for dimensionality using RPCA. In RPCA, the important consideration is to evaluate the contrasts (Linacre, 2014a) between any opposing factors illuminated by RPCA. If the first contrast is much larger than a chance eigenvalue of 2.0, then a further investigation of the items is necessary to discover whether there are off-dimension items that are resulting in threats to the Rasch measurement. An eigenvalue of 2.0 represents the smallest number of items that could represent a second dimension. However, it should be noted that an eigenvalue greater than 2.0 may simply suggest an intensification of the primary dimension, which can be diagnosed by investigation of the fit statistics for the contrasting items in the RPCA.

Additional support for unidimensionality is provided by calculating the ratio of the percent of raw variance explained by the measures (persons and items) to the percent of total variance explained in the first contrast. Ratios exceeding three suggest a strongly unidimensional set of items. In addition to contrast eigenvalues and variance ratios, disattenuated correlations can be calculated for person measures on clusters of items on the instrument. *WINSTEPS* partitions items into three clusters, obtains person measures on each of the three clusters, and reports disattenuated correlations between person measures on each of the three clusters. Correlations approaching 1.0 indicate empirically that the clusters of items are measuring the same thing and that the measure is likely unidimensional.

Preliminary evaluations. Prior to investigating the data for dimensionality using the metrics described above, some preliminary metrics were first reviewed. These are

reported here before discussing details of each domain scale's RPCA. Results of these reviews support moving forward with dimensionality analyses.

Raw variance explained by measures. Linacre (2014a) suggested that a value of raw variance explained by the measures greater than 50% is good and confirms the successful estimation of Rasch measures. For the four domain scales of the SRBCSS-III, the raw variance explained by the measures was good for the four domain scales: 80.4% (mathematics), 73.3% (science), 75.2% (technology), and 80.3% (reading). Moreover, the observed and expected raw variance explained by the measures comported well for the four domain scales, varying by just 0.37% (mathematics), 0.14% (science), 0.40% (technology), and 0.62% (reading). High raw variance explained by the measures and low differences between observed variance explained and expected variance explained provide a first check on the data fit to a model before moving forward in investigating the unidimensionality of the scales.

Indicators of local independence in items. An underlying assumption of Rasch measurement is the notion of local independence (Bond & Fox, 2007). Local independence requires that a participant's response on one item is not dependent on his or her response to another item. This was evaluated using *WINSTEPS* by investigating the standardized residual item correlations, which were obtained via ICORFILE from the output file menu.

High, positive correlations between standardized residuals indicate local item dependency. Except for two positive residual correlations on the mathematics scale (between *understands concepts and processes more easily than other students* and

displays a strong number sense, $r = .11$; and between *solves math problems abstractly* and *displays a strong number sense,* $r = .07$) the residual item correlations were all negative or zero, which suggests that the items reflect local independence. The small positive value of the correlations between the two pairs of math items was likely far too small to present a dependency concern, as values below .30 are generally not considered to indicate a violation of local independence (Christensen, Kreiner, & Mesbah, 2013). However, Christensen et al. added that residual item correlations on instruments of fewer than 20 items may not be as confidently interpreted as those with a large number of items. To overcome this limitation, it is imperative to compare the magnitude of each residual item correlation to the average residual correlation for all items in a set rather than solely on cut-off value criteria. The .11 and .07 correlations between the two pairs of items on the mathematics scale were not particularly larger than the average correlation on the mathematics scale of -.11. Thus, there is strong evidence that the items on the mathematics scale, too, exhibit local independence.

Point-measure correlations. Essential to all analyses is a review of point-measure correlations, which are the correlation of the items with the measure. The point-measure correlations were positive on all retained items for the four domain scales. Point-measure correlations for all retained items on the domain scales are shown in Appendix D.

Mathematics scale dimensionality. Table 11 shows the RPCA data for the 10 retained mathematics items on the field test of the mathematics scale. The total variance explained by the first factor of residuals, 16.3%, was just slightly less than one-fifth as large as the variance explained by the measures, which suggested that the Rasch

dimension is unidimensional. Further evidence of the unidimensionality of the mathematics scale is provided by the first contrast eigenvalue of 1.6, which showed that the unexplained variance did not have the strength of more than two items. Additionally, the disattenuated first contrast person-measure correlations on the item clusters were .99 (item cluster 1-3) and 1.00 (item clusters 1-2 and 2-3), which further supported that the mathematics scale was measuring a single underlying construct.

Table 11

Rasch PCA Results: Mathematics Scale - Retained Items

	Eigenvalue Units	Observed, %	Expected, %
Total raw variance in observations	65.0	100.0	100.0
Raw variance explained by measures	55.0	84.6	84.4
Raw variance explained by persons	49.2	75.7	75.5
Raw variance explained by items	5.8	8.9	8.9
Raw unexplained variance (total)	10.0	15.4	15.6
Unexplained variance in 1st contrast	1.6	2.5 (of unexplained variance)	
		16.3 (of total variance)	

Science scale dimensionality. Table 12 shows the RPCA data for the seven retained science items on the field test of the science scale. The variance explained by the measures was more than three times the total variance explained by the first contrast, which supported that the science scale Rasch dimension was unidimensional. Further supporting the unidimensionality of the science scale was the value of the first contrast

eigenvalue, 1.4, which indicated that the unexplained variance had the strength of fewer than two items. The disattenuated first contrast person-measure correlations on the item clusters were 1.00 (item clusters 1-3 and 1-2) and .99 (item cluster 2-3), which further supported that the science scale had a unidimensional nature.

Table 12

Rasch PCA Results: Science Scale - Retained Items

	Eigenvalue Units	Observed, %	Expected, %
Total raw variance in observations	27.2	100.0	100.0
Raw variance explained by measures	20.2	74.2	74.0
Raw variance explained by persons	15.2	55.8	55.6
Raw variance explained by items	5.0	18.4	18.3
Raw unexplained variance (total)	7.0	25.8	26.0
Unexplained variance 1st contrast	1.4	5.3 (of unexplained variance)	
		20.7 (of total variance)	

Technology scale dimensionality. As with the mathematics and science scales, the technology scale showed empirical evidence of unidimensionality. Table 13 shows the results of the RPCA analysis of the seven-item technology scale. The unexplained variance in the first contrast was less than 2.0 eigenvalue units, and the ratio of the variance explained by the first contrast of residuals was less than one-fourth of the variance explained by the measures. Strong person measure correlations between the first

contrast clusters .97 (cluster 1-3) and 1.00 (clusters 1-2 and 2-3) further suggested a single underlying latent variable was being measured.

Table 13

Rasch PCA Results: Technology Scale - Retained Items

	Eigenvalue Units	Observed, %	Expected, %
Total raw variance in observations	35.2	100.0	100.0
Raw variance explained by measures	28.2	80.1	79.8
Raw variance explained by persons	24.0	68.0	67.8
Raw variance explained by items	4.3	12.1	12.0
Raw unexplained variance (total)	7.0	19.9	20.2
Unexplained variance 1st contrast	1.4	3.8 (of unexplained variance)	
		19.3 (of total variance)	

Reading scale dimensionality. Table 14 shows the results of the RPCA of the reading scale. Similar to the previous three scales, evidence for unidimensionality was provided by the 3.5:1 ratio between the percent of the variance explained by the measures and the percent of total variance explained in the first contrast. As well, empirical support was provided for the unidimensional nature of the scale by both the eigenvalue of the first contrast (1.4) and the strong correlations between persons and measures of .99 (clusters 1-3 and 2-3) and 1.00 (cluster 1-2).

Table 14

Rasch PCA Results: Reading Scale - Retained Items

	Eigenvalue Units	Observed, %	Expected, %
Total raw variance in observations	35.8	100.0	100.0
Raw variance explained by measures	29.8	83.2	82.9
Raw variance explained by persons	24.9	69.5	69.3
Raw variance explained by items	4.9	13.7	13.7
Raw unexplained variance (total)	6.0	16.8	17.1
Unexplained variance 1st contrast	1.4	4.0 (of unexplained variance)	
		24.1 (of total variance)	

Item retention analyses. The SRBCSS-III's authors' 73-item field test conducted for the domain scales was administered to teachers of 726 students in 140 elementary schools (Renzulli et al., 2009). The field test instrument consisted of 20 mathematics items, 24 science items, 9 reading items, and 20 technology items. Using confirmatory factor analysis, the authors determined the best-fitting models to retain 11 mathematics items (although one loaded on more than one factor and was later removed), 7 science items, 6 reading items and 7 technology items.

The next section presents the results of Infit MnSq and Outfit MnSq for the retained items using Rasch analysis, which highlights how well the data accord to the Rasch model. In the analyses, accordance of retained items with the measurement system was determined by the values of Infit MnSq and Outfit MnSq, the mean of which is expected to be near 1.00. Values on items lower than 1.00, model overfit, suggested

redundancy and data that were too predictable, while values on items greater than 1.00, model underfit, reflected noise in the data through unusual response patterns.

Values for Infit MnSq and Outfit MnSq of the items retained on the four domain scales are shown in Tables 15 to 18 (letters or numbers refer to item positions on the scales' general keyforms, Figures 5 to 8).

Table 15

Fit indices: Mathematics Scale

Item	Item Text	MnSq	
		Infit	Outfit
A	uses a variety of representations to explain math concepts	1.20	1.18
C	solves math problems abstractly	1.09	1.04
B	is eager to solve challenging math problems	1.09	1.04
D	has an interest in analyzing the mathematical structure of a problem	1.01	0.94
e	enjoys challenging mathematics puzzles, games, and logic problems	0.96	0.95
c	displays a strong number sense	0.93	0.91
d	understands concepts and processes more easily than other students	0.93	0.89
E	can switch strategies easily, if appropriate or necessary	0.90	0.98
b	organizes data and information	0.89	0.92
a	has creative (unusual and divergent) ways of solving problems	0.82	0.86

Table 16

Fit indices: Reading Scale

Item	Item Text	MnSq	
		Infit	Outfit
1	pursues advanced reading material independently	1.07	1.04
4	applies previously learned literary concepts to new reading experiences	1.02	1.10
8	shows interest in reading other types of reading materials	1.00	1.03
3	focuses on reading for an extended period of time	0.98	0.93
2	eagerly engages in reading-related activities	0.91	0.94
7	demonstrates tenacity when posed with challenging reading	0.90	0.90

Table 17

Fit indices: Science Scale

Item	Item Text	MnSq	
		Infit	Outfit
B	reads about science-related topics in his/her free time	1.17	1.15
A	clearly articulates data interpretation	1.12	1.17
C	expresses interest in science project or research	1.14	1.09
D	is curious about why things are as they are	1.09	1.06
c	demonstrates curiosity about scientific processes	0.89	0.88
b	demonstrates creative thinking about scientific debates or issues	0.79	0.79
a	demonstrates enthusiasm in discussion of scientific topics	0.69	0.71

Table 18

Fit indices: Technology Scale

Item	Item Text	MnSq	
		Infit	Outfit
A	incorporates technology in developing creative products	1.56	1.49
B	spends free time developing technology skills	1.01	1.03
C	eagerly pursues opportunities to use technology	0.98	0.93
D	assists others with technology related problems	0.88	0.92
c	demonstrates more advanced technology skills than other students	0.84	0.85
b	learns new software without formal training	0.83	0.81
a	demonstrates a wide range of technology skills	0.75	0.74

For mathematics, reading, and science, all items comported well to the measurement system, reflecting modest levels of underfit or overfit to the model. Indeed, these items were productive for measurement for the three scales. The fitting of these items to the model can easily be seen in Figures 5 to 7, which graphically show the items as functions of their fit along the measure continuum; indeed, no items reflected misfit relative to others.

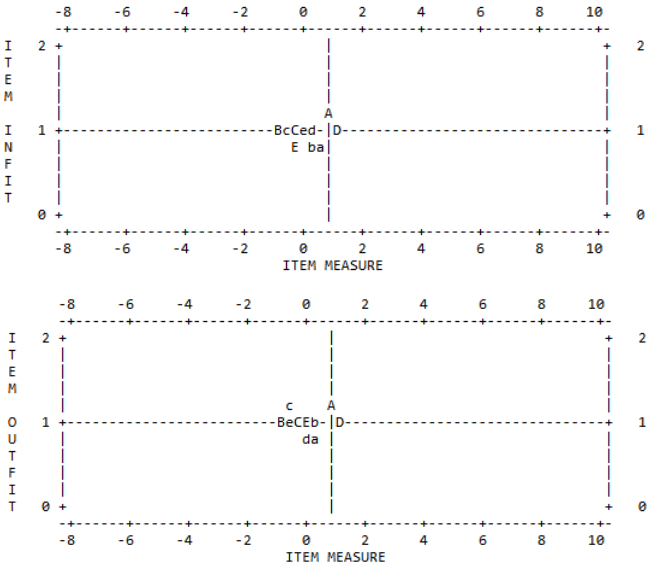

Figure 5. *WINSTEPS* output files of mathematics scale Infit and Outfit MnSq.

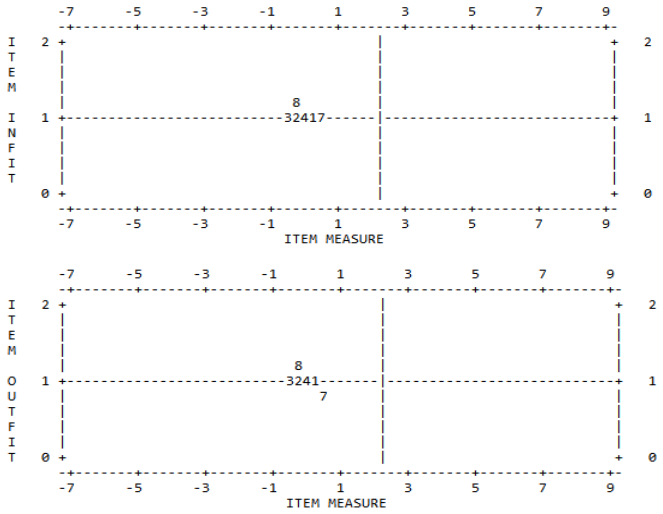

Figure 6. *WINSTEPS* output files of reading scale Infit and Outfit MnSq.

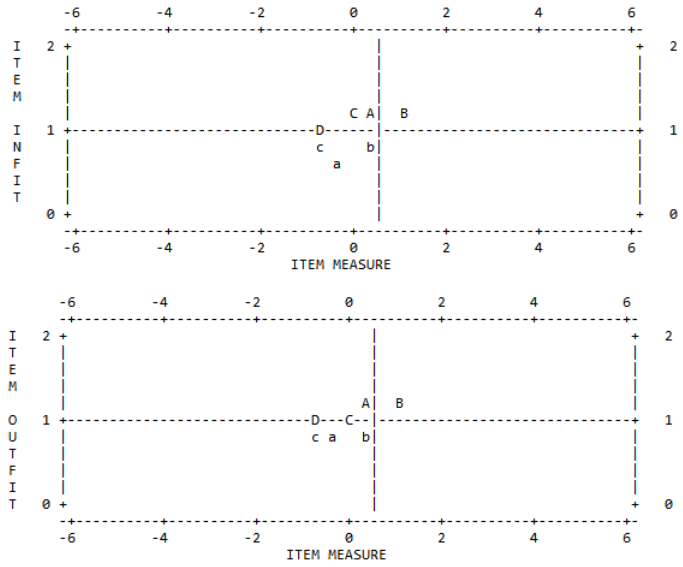

Figure 7. WINSTEPS output files of science scale Infit and Outfit MnSq.

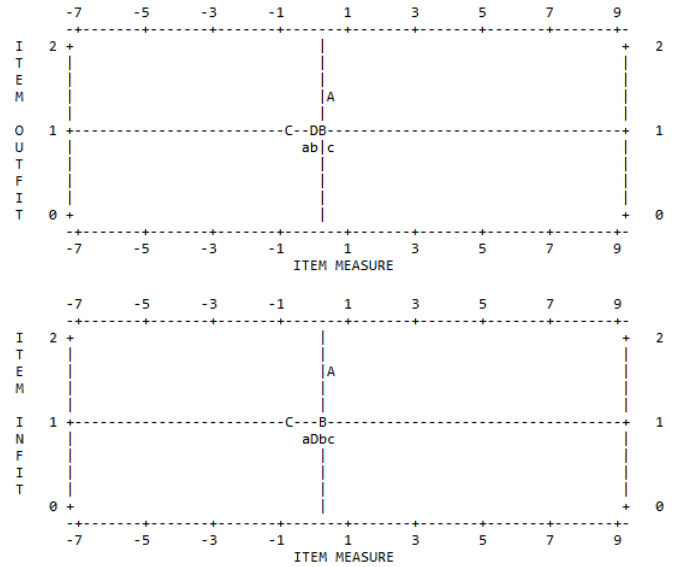

Figure 8. WINSTEPS output files of technology scale Infit and Outfit MnSq.

The technology scale was the only scale to retain an item that continued to misfit at the greater than 1.5 criterion; Figure 8 shows the non-conforming position of the item, Item A. Item A, *The student incorporates technology in developing creative products/assignments/presentations*, showed Infit MnSq of 1.56, indicating that the item reflected a high degree of misfit for in-target students.

Results of Analyses for Research Question 2

In answering the second research question, analyses were performed to evaluate the reliability and evidence of validity for the mathematics, reading, science and technology scales (using the SRBCSS authors' validation field study data) and for the motivation and learning characteristics scales (using the local Grade 3 operational administration data).

The results of the analyses to answer the second research question are presented in the following sections. First, point-measure correlations of items are discussed, followed by person separations and reliabilities. Next, item separations and reliabilities frame a discussion evaluating the item hierarchy as a component of content validity. Finally, overall Infit and Outfit of persons and items are presented, and the section concludes with discussions of the item-person maps for each scale, which revisits the item hierarchy of the scales.

Table 19

Point-Measure Correlation Ranges for High and Low Category-Option Groupings: SRBCSS-III Domain, Motivation, and Learning Characteristics Scales

Scale	Point-Measure Correlation Range	
	Lower Category Group (0,1,2)	Higher Category Group (3,4,5)
Mathematics	-.67 to -.10	-.04 to .58
Reading	-.45 to -.26	-.27 to .73
Science	-.55 to -.08	-.19 to .61
Technology	-.56 to -.13	-.04 to .59
Motivation	-.39 to -.31	-.27 to .67
Learning Characteristics	-.47 to -.35	-.26 to .63

Category point-measure correlations. Each of the scales, as first discussed in the context of Research Question 1 above, exhibited a set of items for which the overall point-measure correlation was large and positive. A point-measure correlation reflects the extent to which person responses on items are in accord with the Rasch requirement that higher category scoring corresponds to the presence of more of the latent variable. Point-measure correlations are shown in Appendix D for all six evaluated scales. In addition to the overall point-measure correlation–i.e., those shown for each scale's items in Appendix D–the item category options should reflect correlations with the measures that support the Rasch measurement model. For each scale, Table 19 shows the range of category option point-measure correlations for the lowest and highest groupings of category options, which shows the accordance of category options' correlations with person measures.

Table 20

Person separation and person and item reliability: SRBCSS-III Domain, Motivation, and Learning Characteristics Scales

Scale	Real Person Separation	Real Person Reliability	Real Item Reliability
Mathematics	6.00	.97	.97
Reading	4.65	.96	.89
Science	3.60	.92	.99
Technology	4.44	.95	.95
Motivation	5.46	.97	.97
Learning Characteristics	6.05	.97	.99

Person separation and person reliability. Person separation can be used to classify students (Linacre, 2014a) rated on instruments such as the SRBCSS-III. A sufficiently large person separation (> 2, person reliability > .80) according to Linacre, Smith (2003) and Bond and Fox (2007) indicates that an instrument is sensitive enough to distinguish between high and low students. This is especially important on an instrument such as the SRBCSS-III, which purports to be able to provide information to teachers in classrooms about the characteristics of their students that suggest higher ability (Renzulli & Smith, 2010). Person reliability expresses the extent to which students estimated at the high range of ability have higher Rasch measures than those with low ability (Linacre). Real person separation and real person and item reliabilities for the six evaluated scales are shown in Table 20.

As Table 20 shows, each of the six scales placed persons reliably on its respective measurement continuum, which suggested that person placement on another instrument measuring the same construct will likely show the same students above or below others as on these scales. The scales operated in such a way as to group students by measure. Person separation indices indicated that persons were successfully positioned along a wide range of the scales' continua. Students were separated into no fewer than about five groups (science; about four separations) and into as many as about seven groups (mathematics, motivation, and learning characteristics; about six separations) by measure. It is important to consider the factors that lead to definitive discrimination of persons: a large sample ability variance, an appropriately-targeted sample, and a low percentage of missing data.

Item reliability and item hierarchy. The real item separation reliabilities indicated that the item set on each of the six evaluated scales creates a well-defined variable and that the items formed a reproducible item hierarchy.

Importantly, however, the item hierarchy must be more than reproducible; it must also represent a conceptually intended order along the latent variable (Linacre, 2014a). This hierarchy can be reviewed through several diagnoses tables in *WINSTEPS*, one of which is the general keyform (also termed the construct keymap), which is Table 2.2 in *WINSTEPS* 3.81.0 (Linacre, 2014b). Keyforms are shown in Figures 9 to 14.

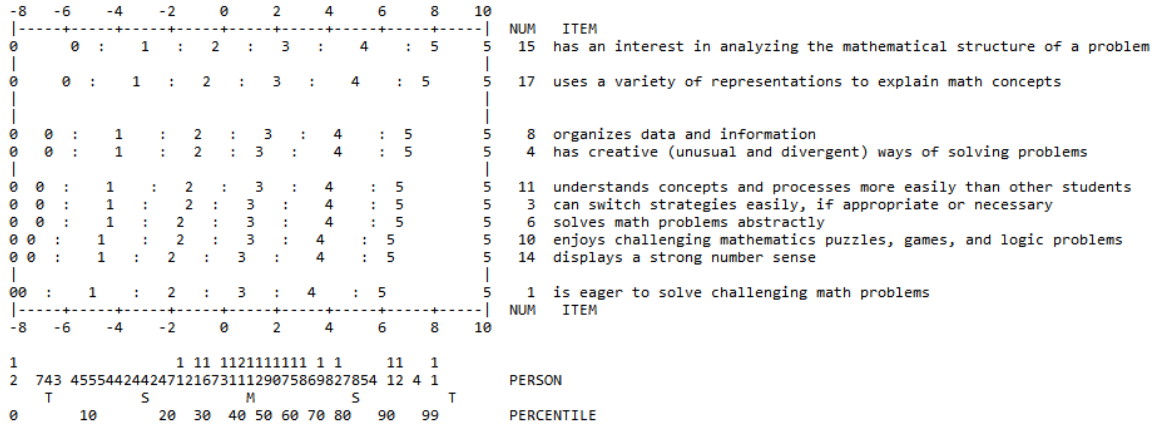

Figure 9. General keyform: mathematics scale.

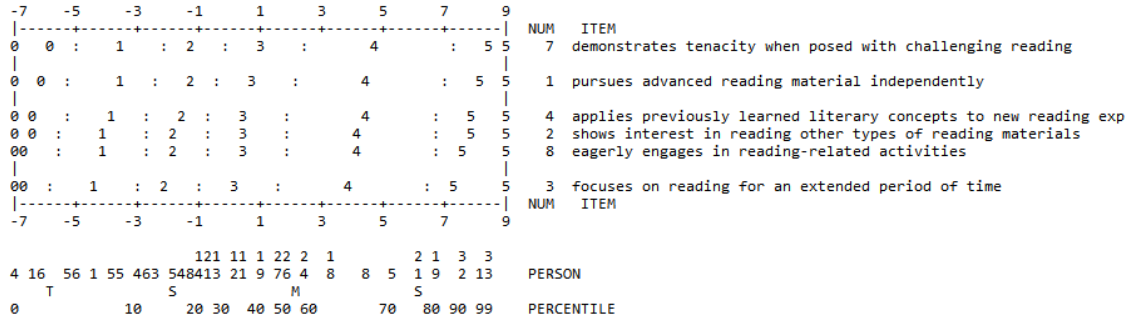

Figure 10. General keyform: reading scale.

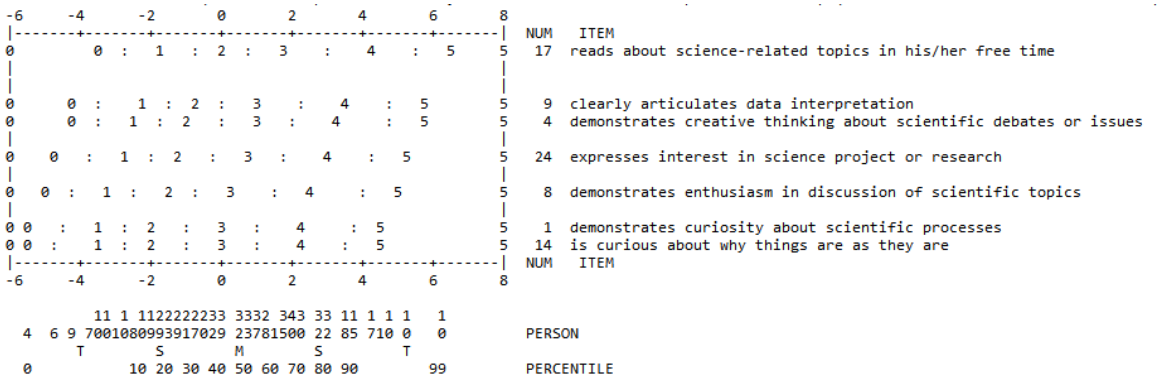

Figure 11. General keyform: science scale.

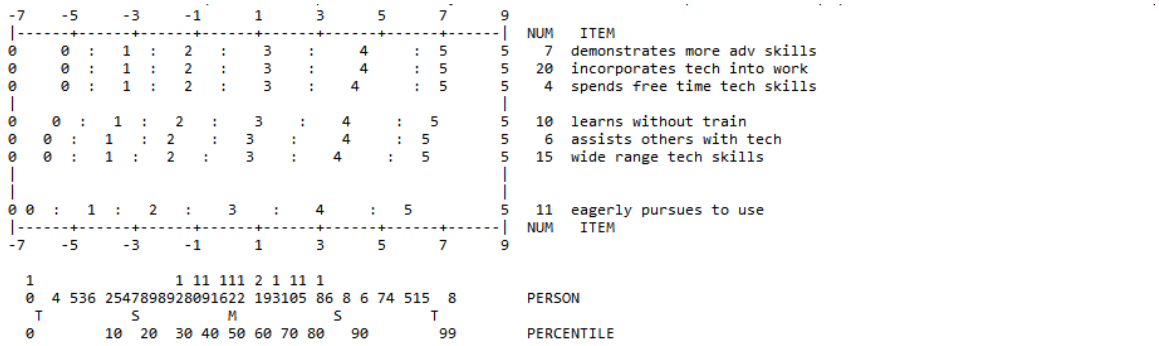

Figure 12. General keyform: technology scale.

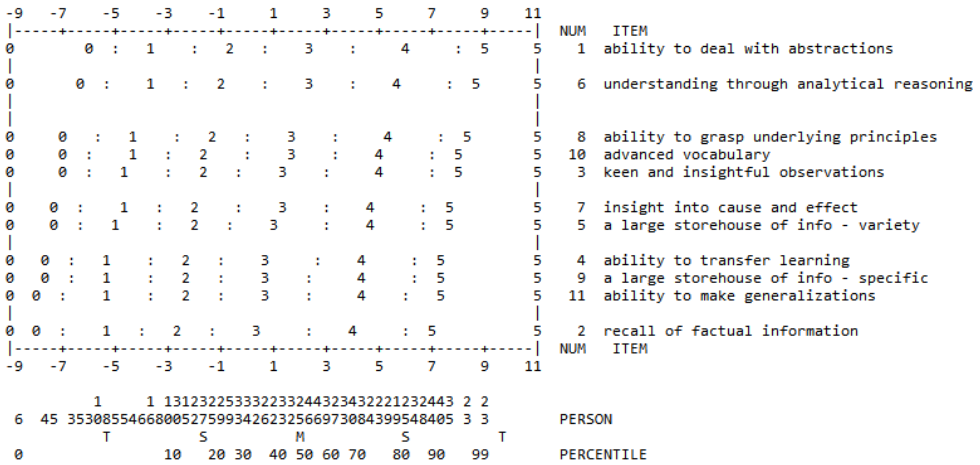

Figure 13. General keyform: learning characteristics scale.

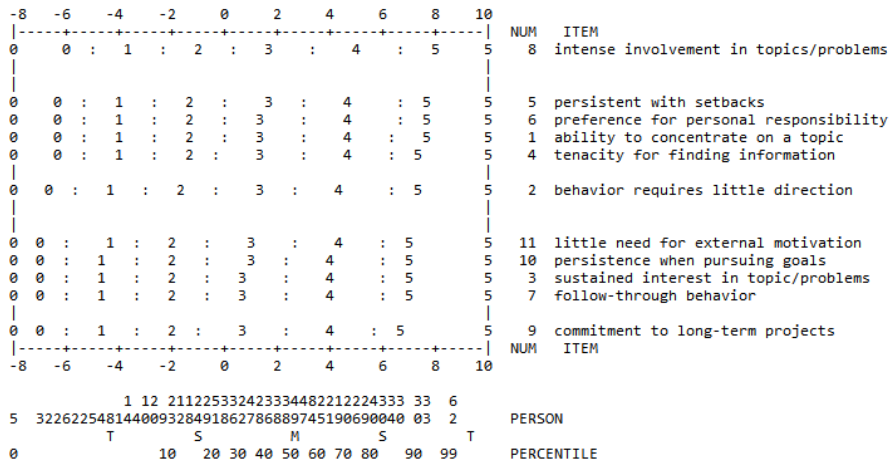

Figure 14. General keyform: motivation scale.

The keyforms represent the item hierarchy for each of the six scales evaluated to answer Research Question 2. In each figure, the most difficult items to endorse at higher option categories are shown at the top, while the items easier to endorse at higher option categories are shown at the bottom. If the alignment of the items in such an order

supports the underlying conceptual nature of the construct in terms of an increasing difficulty to endorse higher option categories for higher-placed items, then there is evidence for content validity for the scales. This will be further explicated in the discussion section of Chapter 5.

Fit indices. Because the Rasch model is a stochastic model rather than a deterministic model, some amount of misfit is tenable and expected (Bond & Fox, 2007; Linacre, 2014a). Callingham and Bond (2006) indicated that as many as 5% of the persons can exhibit misfit without causing concern in the social sciences (although they noted that for a high-stakes situation a 5% misfit rate on *items* must be addressed), while Linacre suggested that up to 10% misfit can be expected and not particularly interfere with Rasch measurement.

Table 21

Mean-Square Summary Fit Statistics: SRBCSS-III Domain, Motivation, and Learning Characteristics Scales

Scale	Person Fit Statistics		Item Fit Statistics	
	Infit MnSq	Outfit MnSq	Infit MnSq	Outfit MnSq
Mathematics	0.98	0.97	0.98	0.97
Reading	0.99	0.99	0.98	0.99
Science	0.99	0.98	0.99	0.98
Technology	0.97	0.96	0.98	0.97
Motivation	0.98	0.97	0.98	0.97
Learning Characteristics	0.98	0.98	0.98	0.98

For the data used to answer Research Question 2, summary fit indices were evaluated for each of the six scales to evaluate the fit of the data to the Rasch model. Table 21 shows that MnSq for both persons and items approached the expected value of 1.0 on all scales, indicating that useful fit of the data to the Rasch model was attained. As discussed earlier, all items except *The student incorporates technology in developing creative products, assignments, or presentations* on the technology scale exhibited good fit, and person misfit was within stochastic expectations at 3.7% (math and learning characteristics), 4.7% (motivation), 5.5% (technology), 6.8% (reading), and 9.1% (science).

Item-person maps. The general item-person map for dichotomies (usually generated as Table 12.2 in *WINSTEPS* 3.81.0) does not serve to address the needs of evaluating the content validity for polytomous scales as it does for dichotomous scales. Instead, the item-person map with polytomous item range (generated as Table 1.4 in *WINSTEPS* 3.81.0) was produced for each of the six scales evaluated for Research Question 2.

The item-person map with polytomous item range shows the distribution of items and persons with items placed at their mean calibrations–the location where ratings in the highest and lowest categories are equally probable–as well as two additional positions: the measure-level at which the probability of being rated in or exceeding the bottom category on the scale is .50 (e.g., left-hand item column in Figure 15) and the measure-level at which the probability of being rated in or below the top category on the scale is .50 (e.g., right-hand item column in Figure 15). Everyone measured between the extreme

thresholds, which are marked by the gray bands on Figures 15 to 20, has a chance of being measured above the bottom category of at least one item and below the top category of at least one item. Thus, the item difficulty of an item covers the distance along the measure from its lowest position in the left-hand item column to its highest position in the right-hand item column. The region between these bands can be evaluated to provide evidence that the range of the construct has been captured by the items on an instrument.

The item-person distribution maps in Figures 15 to 20 show that the items selected on the six scales evaluated for the current study well-covered the range of latent trait of those rated with the scales. There are, however, several students on each scale that far exceeded the highest category on the most difficult item; i.e., they are above the upper gray band. Although there is great certainty that these students, indeed, have the highest amount of the trait being measured, there are no items that discriminated between them at that measure level.

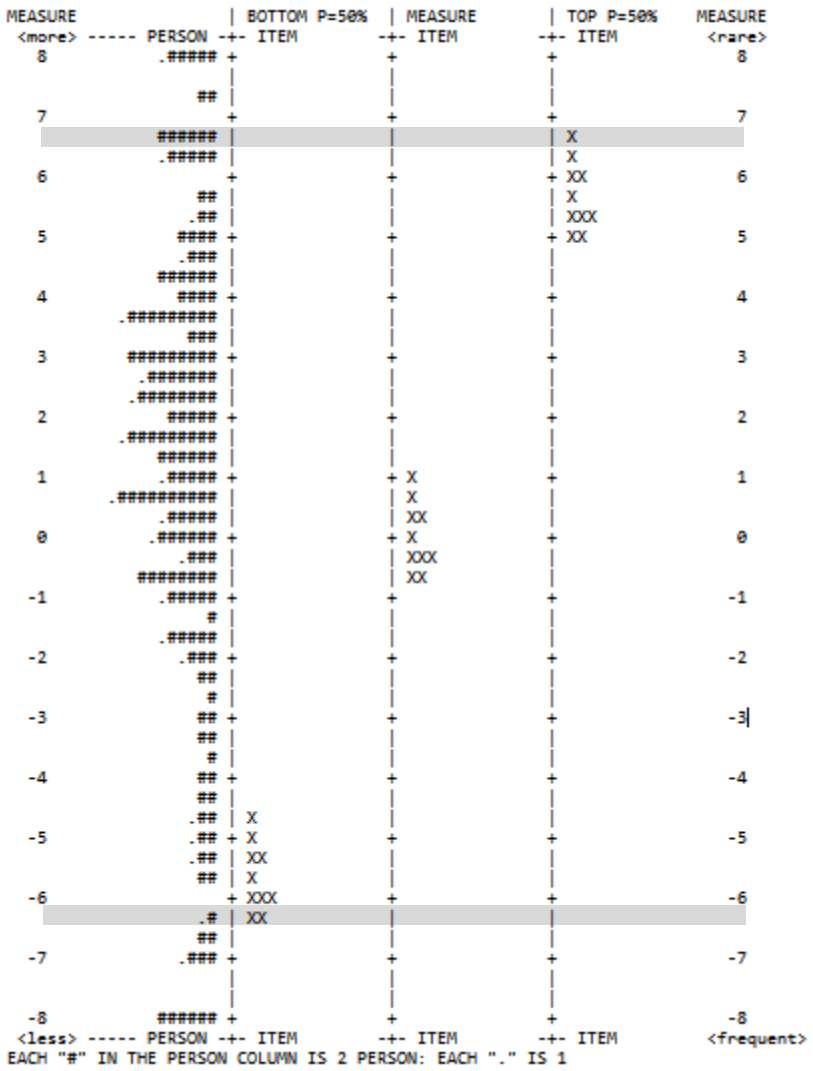

Figure 15. Item-person distribution map: mathematics scale.

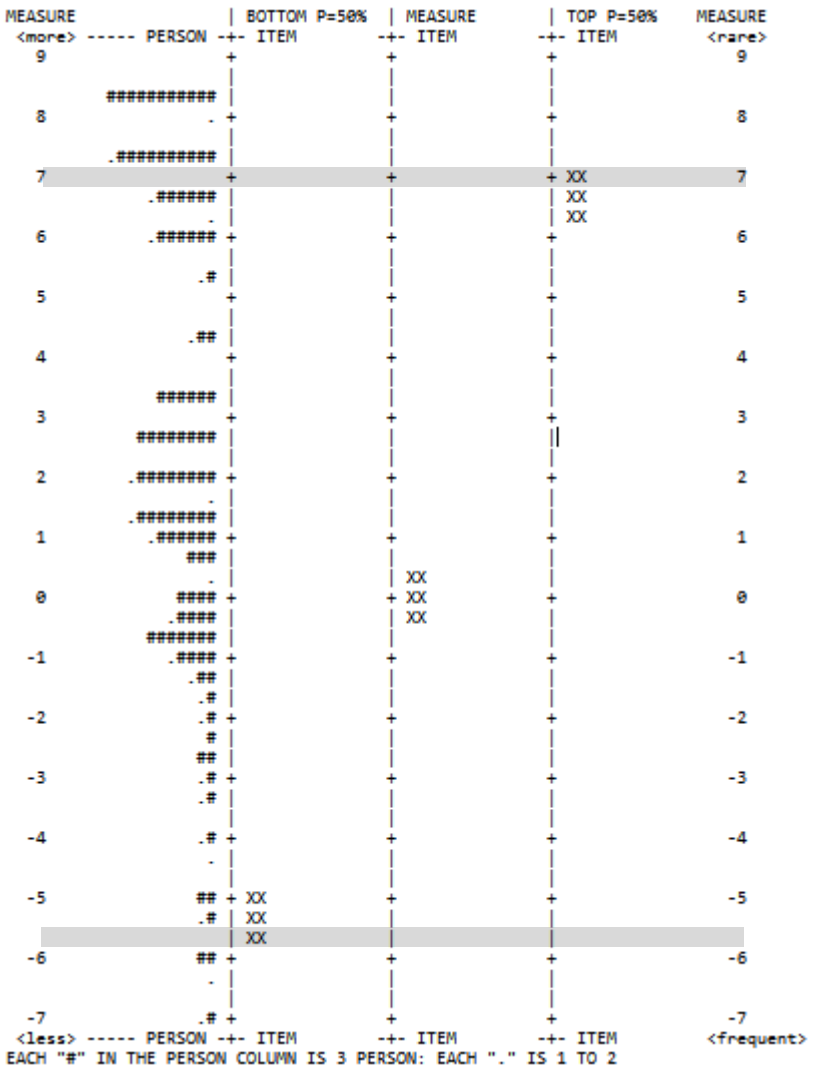

Figure 16. Item-person distribution map: reading scale.

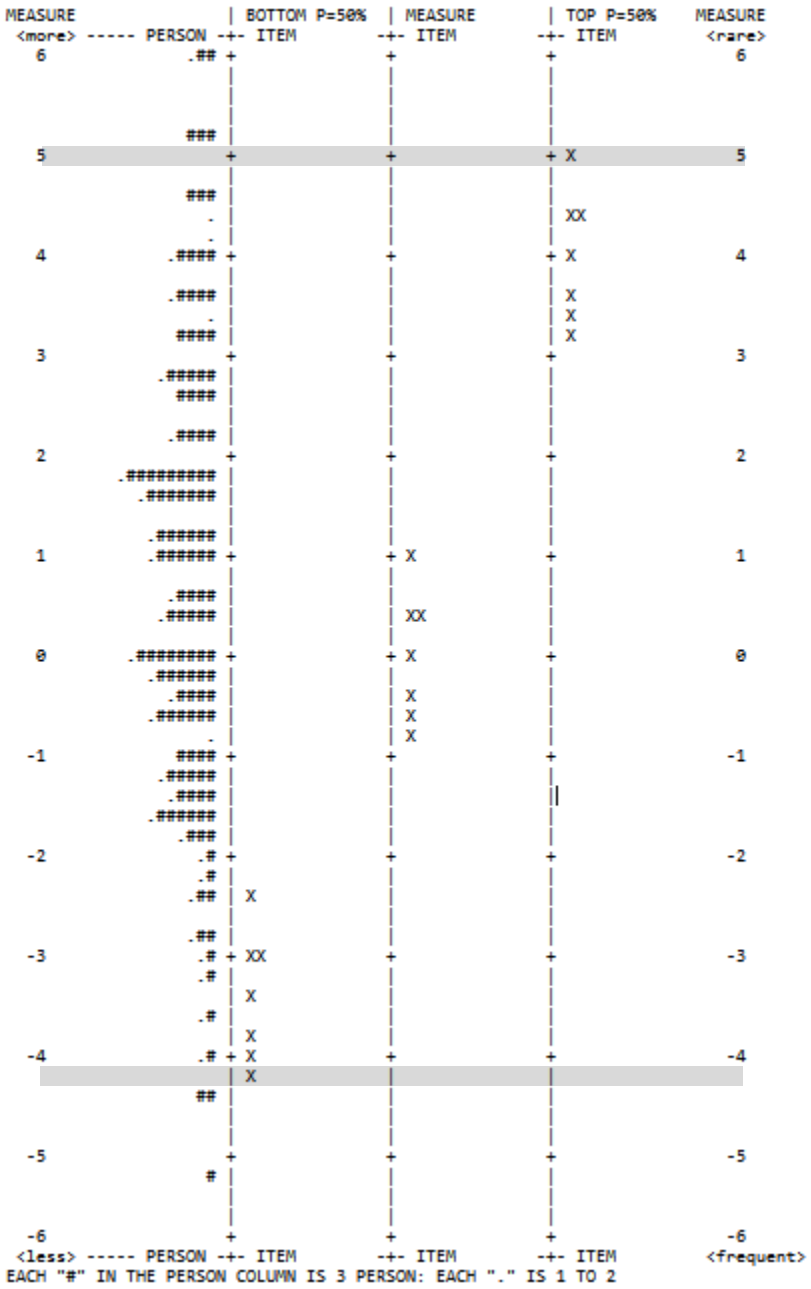

Figure 17. Item-person distribution map: science scale.

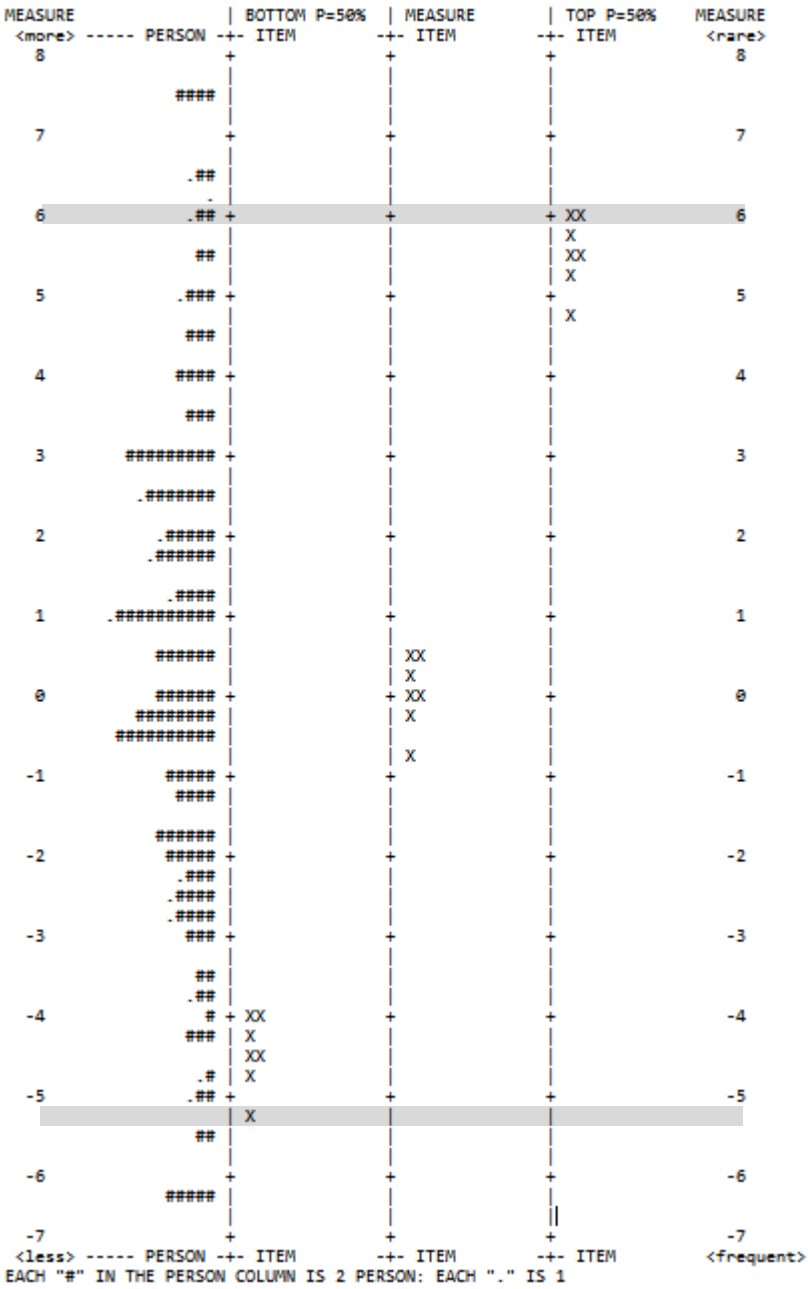

Figure 18. Item-person distribution map: technology scale.

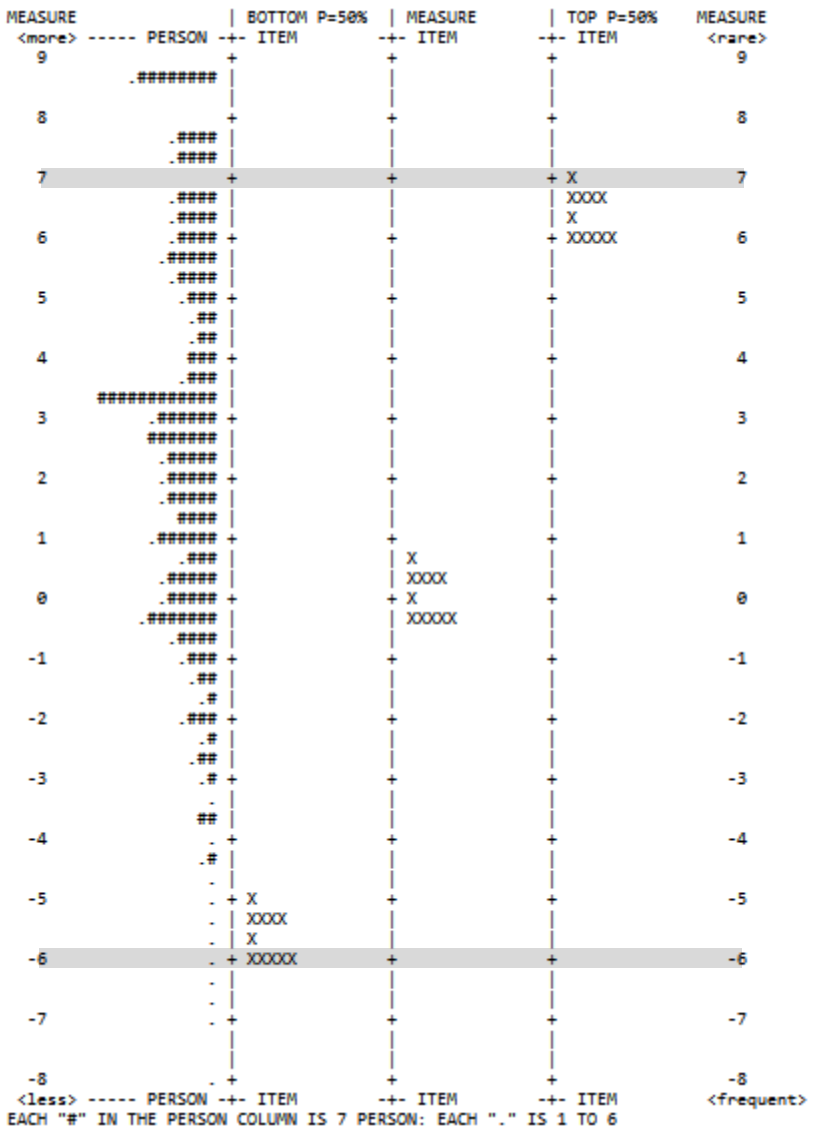

Figure 19. Item-person distribution map: motivation scale.

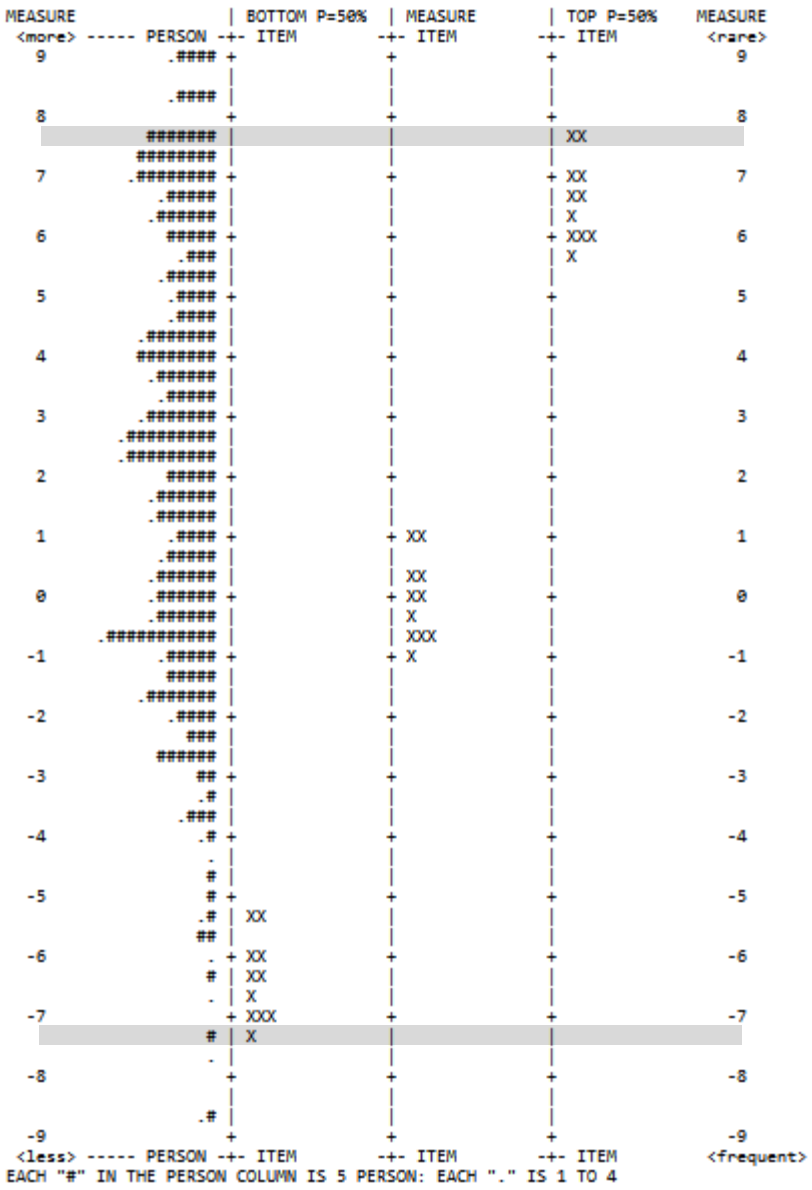

Figure 20. Item-person distribution map: learning characteristics scale.

Results of Analyses for Research Question 3

Substantive validity evidence for a rating scale is shown when empirical evidence reflects that substantively different and meaningful categories were interpreted by respondents. Empirical evidence of substantive validity can be shown through the evaluation of average measures on a scale, coherence between measures and category observations, fit statistics, distributions of categories across items on a scale, and step calibrations.

To answer Research Question 3, the SRBCSS-III authors' validation field test data for the domain scales were used to determine whether optimal category structure exists for the domain scales, evidence of which can support the argument for substantive validity. The next sections present the results of the examination of the domain scales for such evidence.

Category use distributions. Figure 21 shows the distributions of category responses for the four domain scales on the SRBCSS-III. The distributions were generally unimodal distributions, which suggested that there was no aberrant category usage. Importantly, the category distribution on all of the items on each scale was relatively similar for each item on the scale, which created an optimal situation for step calibration.

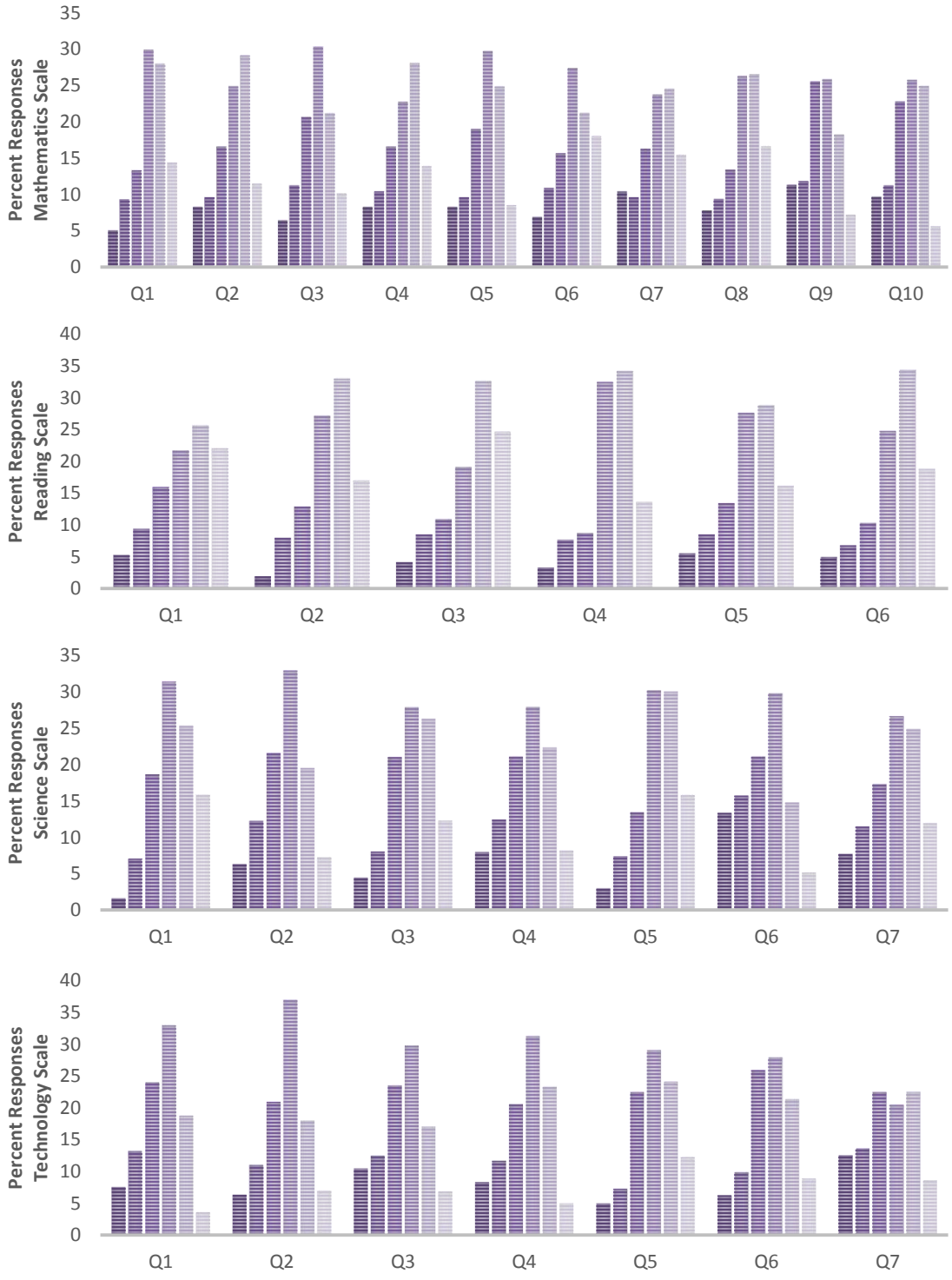

Figure 21. Distribution of responses by category for domain scales items.

Outfit MnSq of categories. As discussed earlier, the Rasch model is a stochastic model that is robust to a uniform level of randomness present throughout the data. Considering data that are too predictable and data that are excessively random (noisy), data that are excessively random are a greater threat to the measurement system (Linacre, 2002). Large, positive values of Outfit MnSq for a category indicate the data may not support useful measurement because large, positive values arise when categories are used in unexpected or idiosyncratic ways. As with item statistics earlier, category Outfit MnSq greater than 1.5 indicates such idiosyncratic use. And, more to the point, values greater than 2.0 indicate there is more unexplained noise than explained noise. Table 22 shows the Outfit MnSq values for the categories on the four SRBCSS-III domain scales. Here, the categories on the scales were being used in a contextually predictable manner, as Outfit MnSq values were very near to 1.0 in all cases. No category usage presented greater than 12% unmodeled noise, which is well within acceptable fit for successful Rasch measurement.

Table 22

Outfit MnSq of Categories SRBCSS-III Domain Scales

Scale	Category Label					
	1 *never*	2 *very rarely*	3 *rarely*	4 *occasionally*	5 *frequently*	6 *always*
Mathematics	0.99	0.91	0.96	1.04	0.98	0.91
Reading	0.97	0.89	0.97	1.05	0.99	1.01
Science	0.88	0.93	0.94	1.12	0.96	0.94
Technology	0.87	0.87	0.92	1.07	1.00	0.98

Average measures and step calibrations. Fundamental to the Rasch model is the notion that observations in higher categories must be produced by persons with higher measures. Thus, average measures by category must advance monotonically moving from lower categories to higher categories owing to the implication that the probability of selecting a higher category increases for respondents with higher measures (Linacre, 2002). Mathematically, conceptualizing the Rasch model as $X_{ni} = \theta_n - (\delta_i + \tau_k)$ where

X_{ni} is the empirical observation when person n encounters item i,

θ_n is the measure of person n on the logit scale,

δ_i is the difficulty of item i, and

τ_k is a threshold indicating the impediment to being observed in category k relative to category $k - 1$,

the average measure relative to item difficulty difference, $\theta_n - \delta_i$, for all persons can be used to evaluate the rating scale across categories. (Because the rating scale structure is the same across all items τ_k is a constant that can be neglected.) Observed average measures relative to item difficulties for each scale are presented in Table 23, and expected average measures relative to item difficulties for each scale are shown in Table 24.

Table 23 shows that the average measures demonstrated monotonicity across all scales measured on the logit scale. Moreover, for the four domain scales the expected values for the average measures in logits shown in Table 24 did not reflect any concerning differences from their observed values. Additionally, for three scales–mathematics, science and technology–there was generally a uniform change in magnitude

between each pair of categories across each scale, which indicated uniform use of the rating scale across its span. For the reading scale there were much larger changes in average measure between category four and category five and between category five and category six. Linacre (2002) suggested that this could be symptomatic of problems with the rating scale for this scale or simply be a factor reflective of the item and sample distributions. At any rate, however, the expected values provide evidence that the increase, although anomalous, does not likely represent a problem in the rating scale use for the reading scale.

Table 23

Average Measures SRBCSS-III Domain Scales; Observed Values

Scale	Category Label					
	1 *never*	2 *very rarely*	3 *rarely*	4 *occasionally*	5 *frequently*	6 *always*
Mathematics	-6.72	-3.83	-0.62	1.42	3.62	6.31
Reading	-5.58	-3.44	-0.88	0.74	4.00	7.51
Science	-3.47	-2.07	-0.65	0.58	2.31	4.24
Technology	-5.14	-3.17	-1.13	0.74	3.08	5.93

Table 24

Average Measures SRBCSS-III Domain Scales; Expected Values in Logits

Scale	Category Label					
	1 *never*	2 *very rarely*	3 *rarely*	4 *occasionally*	5 *frequently*	6 *always*
Mathematics	-6.75	-3.79	-0.64	1.42	3.65	6.26
Reading	-5.62	-3.34	-0.98	0.79	3.98	7.52
Science	-3.45	-2.00	-0.77	0.62	2.32	4.20
Technology	-5.11	-3.08	-1.22	0.76	3.10	5.93

In addition to the average measures increasing monotonically, Table 25 shows that each advance in step difficulties (Andrich thresholds; logits) was large enough (> 1.0) to support an interpretation that movement across the threshold meaningfully indicated successfully overcoming the impediment from category $k - 1$ to category k, and that each advance was small enough (< 5.0) to prevent a loss of precision in measurement. It should be noted that the value of 4.92 between category five and category six on the reading scale was large enough to noticeably diminish the information for students targeted at that location. This can be seen in Figure 22, which is the test information function for the reading scale. However, the greatest loss of information was midway between step calibrations rather than at the threshold boundaries–these students have already cleared the impediment of $k = 5$, which will not prevent their identification owing to their already high measures.

Table 25

Andrich Thresholds SRBCSS-III Domain Scales

Scale	Category Label					
	1 *never*	2 *very rarely*	3 *rarely*	4 *occasionally*	5 *frequently*	6 *always*
Mathematics	none	-5.64	-2.62	0.01	2.57	5.67
Reading	none	-5.27	-2.45	-0.86	1.83	6.75
Science	none	-3.20	-1.97	-0.53	1.69	4.01
Technology	none	-4.43	-2.84	-0.52	2.25	5.55

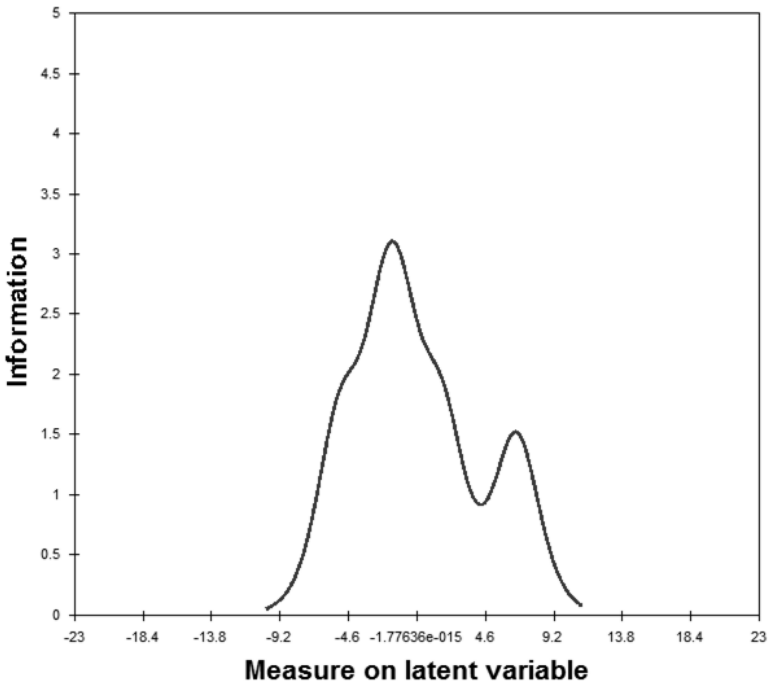

Figure 22. Test information function: reading scale; measure in logits.

Andrich (1996) described that each category on the scale must have a position along the measure at which it is the most probable to be selected. In other words, each category should reflect a modal probability, which was confirmed in Table 25 in that there was no overlap between Andrich thresholds. Linacre (2002) and Andrich asserted that this is essential to the inferential value of an instrument. The category probability curves shown in Figures 23 to 26 illustrate these results visually by highlighting the peaks that represent the modal probability of each category as functions of measure relative to item difficulty in logits.

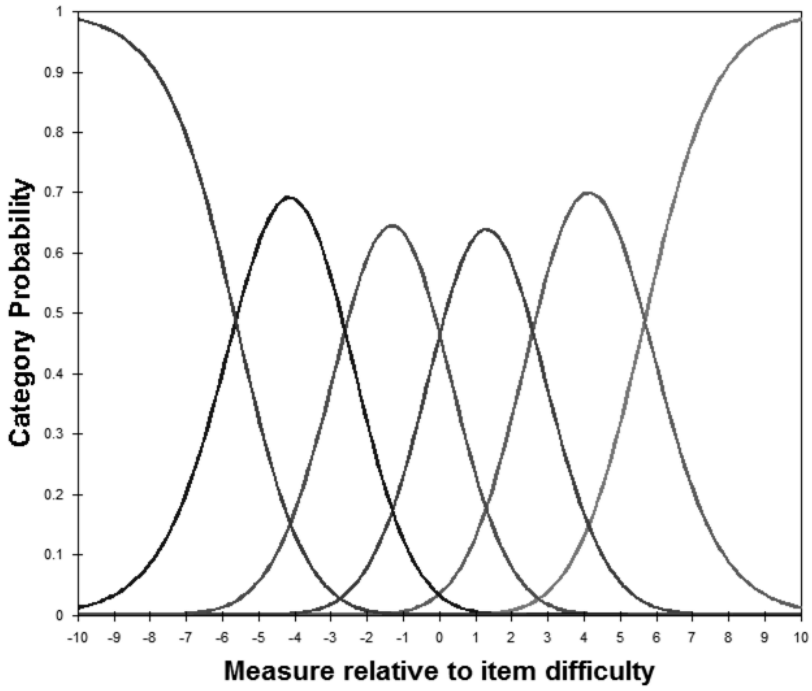

Figure 23. Category probability curves: mathematics scale.

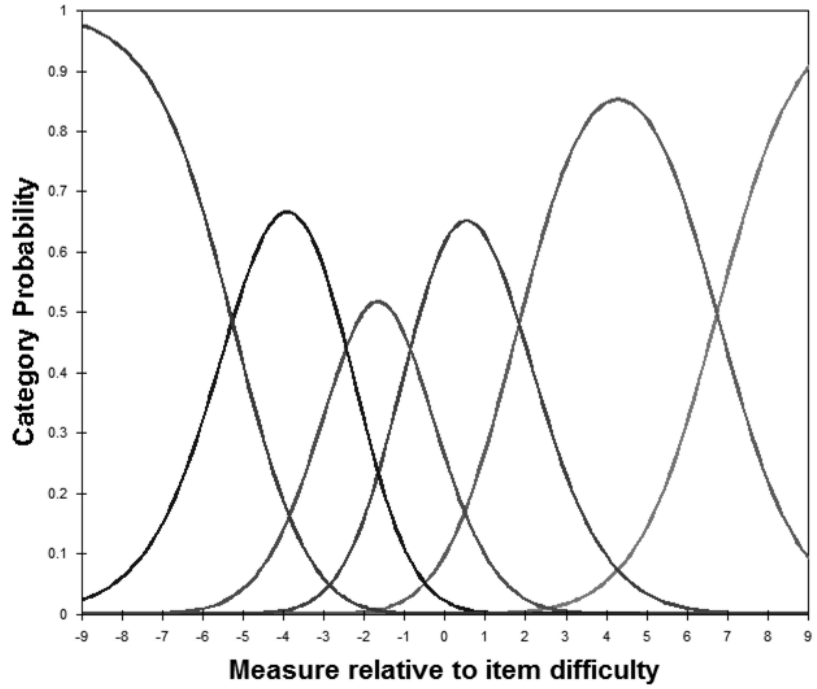

Figure 24. Category probability curves: reading scale.

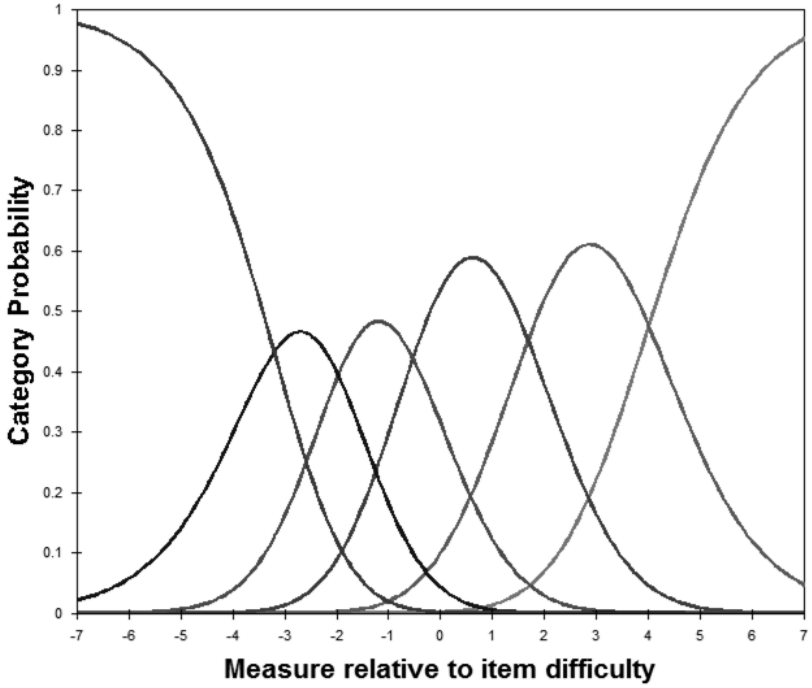

Figure 25. Category probability curves: science scale.

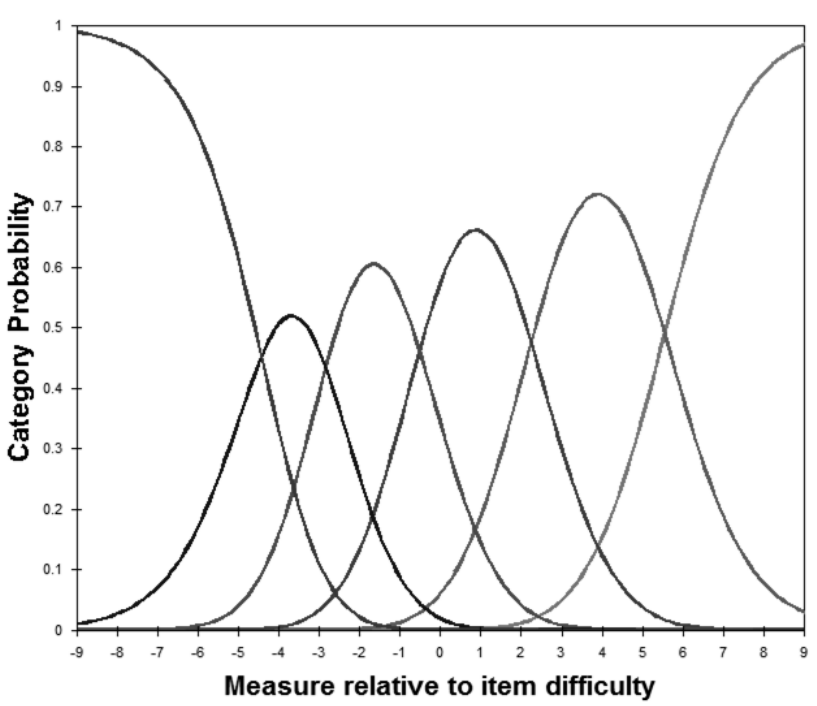

Figure 26. Category probability curves: technology scale.

Coherence measures. Finally, substantive validity evidence can be argued by evaluating coherence measures. The coherence computation expresses the empirical relationship between how well an observed rating implies a person measure and how well a person measure implies an observed rating. For example, Figure 27 shows the item characteristic curve for the mathematics scale. It can be seen in the figure that a student measure of zero showed an expected rating between two and three. Likewise, a measure of zero showed that an average rating of 2.5 would be observed.

.

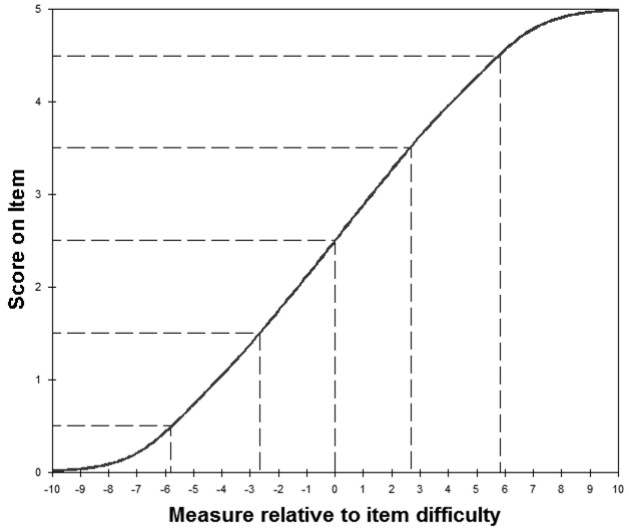

Figure 27. Item characteristic curve for mathematics.

Coherence expresses these empirical relationships as percentages, where coherence greater than 40% (Dimitrov, 2012) suggests a useful level. Coherence percentages are shown in Table 26 and Table 27, and they were generally above the useful level. One category in science (category 1; category implies measure, Table 27) shows a coherence lower than the recommendation of 40%. However, category-implies-measure coherence is often less successful, and values lower than 40% often imply that two categories can be combined.

Table 26

SRBCSS-III Domain Scales Measure-Implies-Category Percentage Coherence

Scale	Category Label					
	1 *never*	2 *very rarely*	3 *rarely*	4 *occasionally*	5 *frequently*	6 *always*
Mathematics	79	62	57	56	62	77
Reading	68	57	42	56	69	73
Science	72	43	44	52	55	74
Technology	80	48	54	58	63	75

Table 27

SRBCSS-III Domain Scales Category-Implies-Measure Percentage Coherence

Scale	Category Label					
	1 *never*	2 *very rarely*	3 *rarely*	4 *occasionally*	5 *frequently*	6 *always*
Mathematics	75	56	56	62	64	66
Reading	59	50	45	64	59	80
Science	38	46	49	55	60	49
Technology	62	49	57	62	63	59

Results of Analyses for Research Question 4

Renzulli et al. (2009) and Jarosewich (2002) have reported on the lack of specific psychometric information at the item and instrument level for scales rating students for identification in gifted programs or talent development settings. Specifically, the authors indicated that review of instruments for performance differences between certain groups of students is scant or does not exist for many widely-used instruments. Analyses such as differential item functioning (DIF) for certain groups have not been completed but should be considered to support the generalizability aspect of validity. In answering Research Question 4, the Grade 4 data from the local data set were used to investigate differential item functioning on the mathematics, reading, motivation, and learning characteristics scales.

For the current study, DIF contrasts on items between groups were used to detect DIF. In DIF contrast, the difference between item difficulty (in logits) for the focal group and the reference group is calculated. *WINSTEPS* calculates a new item difficulty for each item for each group after producing anchor values for person abilities and a rating scale structure for the full sample. It is important to note that differential item functioning arises not just where two groups have varying difficulties with an item at some level; instead, it exists where two groups have varying difficulties with an item and the two groups have been matched on overall ability. An acceptable interpretation (Linacre, 2014a) of the absolute values of DIF contrasts places items in a *negligible* DIF category (|DIF contrast| < 0.43), a *slight-moderate* DIF category (0.43 ≤ |DIF contrast| < 0.64), or a *moderate-large* DIF category (|DIF contrast| ≥ 0.64).

DIF was calculated using three focal groups, female, minority, and economically disadvantaged (ED), compared to three reference groups, male, Caucasian, and not economically disadvantaged, respectively. The results of each are shown in Tables 28 to 31. A negative value indicates that the item was more difficult for the focal group, while a positive value indicates that the item was more difficult for the reference group.

Table 28

DIF Contrast Values: Mathematics Scale

Item	Focal Group		
	Female	Minority	ED
uses a variety of representations to explain math concepts	0.23	-0.04	-0.27
solves math problems abstractly	-0.39	-0.04	0.17
is eager to solve challenging math problems	0.13	0.20	-0.14
has an interest in analyzing the mathematical structure of a problem	-0.22	-0.21	0.00
enjoys challenging mathematics puzzles, games, and logic problems	0.05	0.00	0.23
displays a strong number sense	0.00	0.00	0.17
understands concepts and processes more easily than other students	-0.12	0.02	0.09
can switch strategies easily, if appropriate or necessary	0.19	0.06	-0.53*
organizes data and information	0.27	-0.14	0.33
has creative (unusual and divergent) ways of solving problems	-0.11	0.17	0.00

Note: DIF items marked with an asterisk.

Table 29

DIF Contrast Values: Reading Scale

Item	Focal Group		
	Female	Minority	ED
pursues advanced reading material independently	-0.08	0.11	0.00
applies previously learned literary concepts to new reading experiences	-0.35	0.10	0.00
shows interest in reading other types of reading materials	-0.37	-0.30	-0.05
focuses on reading for an extended period of time	0.58*	-0.16	0.32
eagerly engages in reading-related activities	0.42	0.03	-0.30
demonstrates tenacity when posed with challenging reading	-0.17	0.20	0.00

Note: DIF items marked with an asterisk.

Table 30

DIF Contrast Values: Motivation Scale

Item	Focal Group		
	Female	Minority	ED
the ability to concentrate intently on a topic for a long period of time	0.29	-0.16	0.13
behavior that requires little direction from teachers	0.63*	0.08	0.00
sustained interest in certain topics or problems	-0.52*	-0.04	0.00
tenacity for finding out information on topics of interest	-0.47*	-0.36	0.04
persistent work on tasks even when setbacks occur	-0.18	-0.19	0.00
a preference for situations in which he or she can take personal responsibility for the outcomes of his or her efforts	-0.35	-0.10	0.00
follow-through behavior when interested in a topic or problem	0.38	0.49*	-0.28
intense involvement in certain topics or problems	-0.34	0.00	0.03
a commitment to long-term projects when interested in a topic	0.29	-0.04	0.19
persistence when pursuing goals	0.14	0.16	-0.14
little need for external motivation to follow through in work that is initially exciting	0.21	0.21	-0.10

Note: DIF items marked with an asterisk.

Table 31

DIF Contrast Values: Learning Characteristics Scale

Item	Focal Group		
	Female	Minority	ED
advanced vocabulary for his or her age or grade level	0.29	-0.06	0.00
the ability to make generalizations about events, people, and things	0.16	-0.03	-0.12
a large storehouse of information about a specific topic	-0.43*	-0.05	0.21
the ability to grasp underlying principles	-0.05	-0.27	-0.03
insight into cause and effect relationships	0.26	-0.13	0.40
an understanding of complicated material through analytical reasoning ability	-0.10	-0.18	0.04
a large storehouse of information about a variety of topics	-0.55*	0.21	0.05
the ability to deal with abstractions	0.19	0.32	-0.16
recall of factual information	0.08	-0.08	-0.23
keen and insightful observations	0.12	0.27	-0.14
the ability to transfer learning from one situation to another	0.08	0.03	-0.04

Note: DIF items marked with an asterisk.

Seven items exhibited a slight to moderate DIF, while no items exhibited DIF considered large by the accepted limits presented above. However, it is the case with Rasch analyses that testing such as DIF tests should be reviewed in the context of purpose and practical use (Linacre, 2014a). Statistical tests such as *t*-tests and χ^2 tests often fail in Rasch measurement, which leads to Type I errors. This implies that substantive meaningfulness should be considered alongside statistical testing when using Rasch analysis. Further discussion of the meaningfulness of these results and their implications for the scales' use will be developed in Chapter 5.

CHAPTER FIVE: DISCUSSION AND CONCLUSIONS

As conceptions of giftedness continue to embrace talent development models, instruments that describe students in terms of observed characteristics related to both potential and ability are replacing traditional metrics for identifying giftedness such as IQ or achievement testing (Olszewski-Kublius, 2009; Renzulli & Smith, 2010; Robinson et al., 2007). Giftedness is no longer considered solely within the domain of intellect (Callahan et al., 1995). Subotnik et al.'s (2011) conceptualization, for example, positioned interest and commitment as essential, while Moon and Brighton's 2008 study of the characteristics viewed by educators as representing giftedness highlighted the desire to understand deeply, the transfer of learning, possession of unusual interests, and a faster pace of learning. Sternberg (2001) identified features of students that reflect the development of expertise: motivation, ability to challenge oneself, possession of analytical and evaluative skills, creativity, and readily capable of relating old and new information. Renzulli (1978, 1998) incorporated above average general and specific ability, creativity and task commitment to describe students with potential for placement in programs for gifted students or settings that focus on talent development. As such research illustrates, the notion that giftedness can only be characterized by an IQ-based metric has been changing over the last 30 years as talent development models have emerged.

With changing conceptions of giftedness, instruments that inform educators about the location of students along a continuum of behaviors in the talent development paradigm are needed to ensure that introduction to talent development programming is provided at a developmentally appropriate time and at the right intensity (Renzulli, 2012). For example, Renzulli's (1978, 1998) enrichment model as talent development begins with extended whole-class instructional experiences, which are later followed by interest-driven research and skills-based instruction. The model initially provides opportunities for students to participate in exploration activities and later moves students toward engagement in opportunities that will allow them to display observable attributes along the continuum of talent development. Olszewski-Kubilius (2009) discussed this as creating the optimal environment in which a child's talent can be nurtured. Instruments that can precisely locate students on a continuum of observable behaviors can be used to well-inform teachers about the most appropriate entry points for their students. The SRBCSS-III scales evaluated in the current study possessed the psychometric properties to support such use.

The purpose of this study was to evaluate the validity, characteristics of reliability, item selection, category structure and differential item functioning across student groups of several scales on the SRBCSS-III. This study showed evidence to support the content, structural, substantive, and generalizability validity of inferences made when using the SRBCSS-III, which assures that their use in identification for talent development programming is both justifiable and meaningful. Moreover, this study found evidence in the context of a Rasch measurement model that the scales possessed the

properties of useful measurement described by Bond & Fox (2002, 2007), which include sensitivity to a developmental order and a capability to show developmental distances along a scale.

The discussion that follows first contextualizes the data obtained through the Rasch analyses around four aspects of validity: content validity, structural validity, substantive validity, and generalizability validity. Next, the reliability of item and person measures is discussed. Later, the conclusion places the findings within the perspectives of both the emerging conceptualization of giftedness as talent development and the role of Rasch measurement in constructing measures. Finally, directions for further study are presented.

Validity Evidence

Messick (1989, 1995a) asserted that validity is fundamentally an evaluation argument that leads to a judgment of the degree to which support for interpretation has been substantiated. The argument should be based on a theoretical rationale and a clear understanding of relevant frameworks. In the current study, the theoretical rationale underlying the SRBCSS-III rating scales was the conception of giftedness as talent development, while the framework around which the validity argument was evaluated was the Rasch rating scale model. In building the argument for the current study, four aspects of validity were evaluated using Rasch measurement analyses. The evidence presented in the sections that follow is shown under headings that best illustrate how the empirical information supports a particular aspect of validity. However, it is important to note that the model of evaluating evidence for validity is a unified model (Messick, 1989)

in which multiple sources of evidence support conclusions about the inferences that can be made from an instrument. The presentation below is organized under headings that could very easily overlap, and, in fact, work together to build overall validity evidence.

Substantive validity. The theoretical rationale underlying an instrument is supported where it can be shown that substantive validity evidence is present. The substantive aspect of validity augments the content aspect to the extent that it shows that domain processes and performance regularities (Dimitrov, 2012) comport to theories around which an instrument has been designed. For polytomous Rasch analyses, item hierarchies, response distributions, fit analyses, and category functioning reviews are evaluated to provide evidence for the substantive aspect of validity. The substantive aspect of validity according to Wolfe and Smith (2007b) evaluates "[T]he degree to which the responses…are consistent with the intended cognitive processes around which the [instrument was] developed" (p. 209).

In the current study, item fit statistics for the retained items were generally within stochastic expectations on the content scales, with only one item showing misfit for in-target students. Review of the item category endorsements for the item–*The student incorporates technology in developing creative products/assignments/presentations* found on the technology scale–suggested that there may have been a problem with the forced selection for non-observable behaviors: Are students provided the opportunities to incorporate technology into products? Although the mean of combined *never* and *very rarely* endorsements was 18% for all other items, the percentage of endorsements in these two categories for this item was 27%, or 50% greater than for other items on average.

Additional evidence for this problem can be found where category three, *occasionally*, unexpectedly showed especially large positive category point-measure correlations for some items on the content scales. This was possibly due to the fact that the category option to the left, *rarely*, has an undesirable meaning while *occasionally* has a desirable meaning, and teachers selected the most desirable option for non-observed behaviors for students otherwise rated in higher categories on remaining items. That is, while category three generally did not correlate positively with the highest scoring students, on items involving characteristics less likely to be observed in the classroom the category was positively-correlated with higher measures.

Notwithstanding the potential problem relating to the unobservable behaviors for some items, however, at the category-option level both empirical observations and model expectations for category measures on items on the SRBCSS-III were in accord with one another, supporting the argument for substantive validity. Response distributions were found to be generally similar across not just the items but also across scales, and the scales exhibited generally unimodal distributions, which supported successful step calibration.

Dai and Chen (2013) discussed specific behaviors and abilities related to higher-order cognitive tasks that are required for talent development, which appeared at the top of the item hierarchies on the SRBCSS-III scales. Such hierarchies arise where there is evidence that items associated with aspects of the construct that are conceptually viewed as more difficult are, in fact, aligned with higher measures on the Rasch scale. The current study showed that an item hierarchy existed for the SRBCSS-III that comported

well with the conceptualizations of giftedness as talent development. Located at the top of the SRBCSS-III's scale hierarchies, terms such as 'analyzing,' 'explain,' and 'organizes' on the math scale; 'challenging' and 'independently' on the reading scale; 'interpretation' and 'creative' on the science scale; 'advanced' and allusions to application on the technology scale; 'persistent,' 'intense,' and 'tenacity' on the motivation scale; and 'abstractions,' 'analytical,' and 'advanced' on the learning characteristics scale underlie an item hierarchy consistent with the contemporary conceptualizations of talent development. In addition, large item separation indices with high reliabilities verified the successful placement of items along the item hierarchy for the content scales in the current study. Such conceptually- and empirically-confirmed hierarchical representations of construct-important concepts provided additional evidence for substantive validity for the SRBCSS-III.

Average measures on the SRBCSS-III increased monotonically for students rated in lower categories compared to those rated in higher categories, a pattern that is fundamental to drawing valid inferences from measures. Moreover, the average measures across category options on the scales increased uniformly overall, which indicated successful use of the rating scale categories by respondents. Step difficulties on the SRBCSS-III's scales showed that movement across thresholds reflected a meaningful advancement across impediments to higher category options. Across all scales, step difficulties were wider than 1.0 logit and narrower than 5.0 logits, which provided precision to the measures.

Adding to the substantive validity argument, category probability curves showed a modal probability for each category as a function of measure relative to item difficulty. The presence of a distinct probability at which each category was the most likely is essential to the inferential value of the instrument (Andrich, 1996; Linacre, 2002). As well, coherence measures showed empirically the relationships between how well ratings successfully implied person measures and how well person measures successfully implied ratings.

In summary, evidence of substantive validity in the context of Rasch measurement was found for the SRBCSS-III. Fit statistics and point-measure correlations were found to be in accord with expectations, and response distributions supported a successful step calibration. Moreover, the step calibrations showed a meaningful advancement toward higher category options without loss of precision. Importantly, an item hierarchy clearly emerged that is consistent with the conceptualizations of talent development discussed in Chapter 2.

Structural validity. Structural validity relates to the measurement consistency of an instrument as it is used to assess a construct domain. Evidence for structural validity using Rasch analysis can be shown through the evaluation of the unidimensionality of the scales and the local independence of items. Unidimensionality evaluation assures that an instrument is measuring a single trait or ability, while local independence of items ensures that linkages between questions do not confound item difficulty and person ability estimates (Dimitrov, 2012; Osterlind, 2010).

The results of the RPCA showed that there was an underlying unidimensional nature to the domain scales on the SRBCSS-III researched in the current study. The raw variance explained by the measures was greater than 73% on all scales, and the ratio between the raw variance explained by the measures and the percent of total variance explained in the first contrast exceeded the widely-accepted minimum 3:1 ratio for all scales. Additionally, disattenuated first contrast person-measure correlations calculated by clustering through RPCA were greater than .97 for the all scales and exceeded .99 in 75% of the cluster comparisons.

Standardized residual item correlations showed additional structural validity evidence. Other than the standardized residual item correlations on two pairs of items from the mathematics scale (between *understands concepts and processes more easily than other students* and *displays a strong number sense*; and between *solves math problems abstractly* and *displays a strong number sense*) the residual correlations were negative, indicating local independence that supported a conclusion of the unidimensional nature of the scales. The two pairs that showed slight positive residual correlations had values far too small to result in a suspicion of item dependence.

Overall, Rasch measurement showed evidence for the structural validity of the SRBCSS-III. The RPCA results showed that the scales were unidimensional. Additional evidence for unidimensionality was shown by evaluation of residual correlations. Taken together, these indicated that a single underlying construct was measured.

Content validity. Evidence for content validity is provided by the information that builds an argument for content relevance, content representativeness and item technical quality (Dimitrov, 2012). Evidence for content validity of the SRBCSS-III was shown by several metrics developed in the Rasch analyses.

The results of the Rasch analyses showed that item point-measure correlations between the items and the person measures were in accord with the requirement that higher category scoring corresponds to the presence of more of the latent variable. For all retained items on the four domain scales, the item point-measure correlations were positive, indicating that the items were oriented in the same direction as the measure. As well, category point-measure correlations were also in line with the measures. In other words, for each item, lower category options (0, 1, and 2) showed more-negative point-measure correlations, while more-positive point-measure correlations were seen with higher category options (3, 4 and 5).

Item representativeness was shown in Rasch analyses through the construction of item-person maps, which in the current study showed that the distribution of items was relatively well-spread along the range of person measures. Only a few number of students were located beyond the explicitly marked range of item coverage on the scales. Considering the range of persons and items on the item-person maps and the point-measure correlations of items and categories, strong evidence was found for content validity of the SRBCSS-III.

Generalizability validity. Renzulli and Smith (2010) asserted with the release of the SRBCSS-III the need to evaluate its scales in terms of the demographic characteristics of the students for whom scales will be completed. The current study evaluated the items for differential item functioning (DIF) with respect to sex (male as reference group), ethnicity (Caucasian as reference group), and socioeconomic status (non-economically disadvantaged as reference group). For the evaluated scales here, seven items exhibited slight to moderate DIF using the method of DIF contrast, and none of the items showed large DIF. For several reasons, it is unlikely that the scales need revision given the slight DIF observed–even if similar such DIF were to be found in additional research (Linacre, 2014a):

- recommended levels used in the DIF contrast were strict and are usually used for the case of educational assessments–they can easily be relaxed for other types of instruments;
- DIF contrast values were not large, and they pointed in both directions for the focal and references groups; and
- no single scale represented a large number of DIF items; Rasch measurement is robust with respect to independent item discrepancies.

Reliability

In the context of Rasch measurement reliability represents the reproducibility of person measures or item measures (Bond & Fox, 2007; Linacre, 2014a; Smith, 2003). On the SRBCSS-III scales researched in the current study, the scales showed person reliabilities ranging from .92 to .97 and item reliabilities ranging from .89 to .99. Such

high person reliabilities indicated two characteristics of the scales: a) persons rated high on a scale had a high probability of indeed having a higher measure of the behavioral attributes on the scale and b) persons were successfully placed into groups of varying performance levels. Importantly, this showed that students were clearly placed along a continuum of the measure rather than dichotomized into just high or low groups. Such a continuous measure provides opportunities for programming decisions to be based on a wide variety of ability levels to ensure that students can be exposed to experiences tailored to their developmental level. In addition, the high item reliabilities indicated the replicable position of items in terms of item difficulty; thus, the item reliabilities confirmed the item hierarchies of the scales and add evidence for the substantive aspect of construct validity.

Discussion

The evidence for validity, reliability and essentially absent differential item functioning using Rasch analysis informs the use of the SRBCSS-III in several ways. The current study provided support for the continued use of the scales, and the study confirmed that in many ways the scales can provide educators information for identification within the context of talent development models. Furthermore, the fit of the data to the Rasch model highlights the properties of the scales in terms of the theoretical rationale used to develop the instrument. Together, these ideas are discussed in the sections below in terms of the SRBCSS-III's use in schools and in terms of further development of scales in a changing environment of gifted education as talent development.

Implications for identification. The scales of the SRBCSS-III evaluated in the current study showed evidence that the theoretical rationale of the talent development model was realized in their development. Item hierarchies and the well-functioning categories supported the placement of students along a continuum of progressively-demanding cognitive behaviors. Consistent with Subotnik et al. (2011), evidence from the current study underscored the capacity of the SRBCSS-III to identify the relative nature of a student's behaviors on a continuum from ordinary to extraordinary ability. In terms of nonintellective traits, the scales reliably identified students that show an increasing amount of potential, which cannot be obtained through the administration of achievement tests. Academic achievement often tends to remain constant; identification using instruments modeled on creativity, motivation, and commitment highlights the "contextual, situational, and temporal" (Renzulli, 2012, p. 153) nature of characteristics shown to be essential in the talent development paradigm.

Implications for the classroom. In practical terms, many of the characteristics that are implicit or explicit in curricular models built around college- and career-ready standards are similar to those that have been at the foundation of talent development for many years (Renzulli, 2012). Using instruments such as the SRBCSS-III can be informative for teachers in that the resulting location of students on the continuum can be used to create environments that support identified potential. In doing so, teachers can be better-equipped to ensure that the classroom environment provides opportunities to tap into students' areas of strength to support both talent development as well as the curriculum that is becoming more dependent on skill acquisition and demonstration

rather than dependent on preselected and standardized lessons. Using the information from the SRBCSS-III to support both talent development as well as curriculum align well with contemporary conceptualizations of giftedness as talent development.

Implications in terms of measurement and scoring. Showing that the data comported to the Rasch model highlighted properties of the SRBCSS-III that provide support for the use of the instrument in several ways. First, data fit in terms of the Rasch rating scale model confirmed the presence of the property of fundamental measurement, or that the magnitudes of students (measures) or items (difficulties) along the continuum reflected equal-interval spacing on the Rasch scale and a quantifiable and meaningful difference between students or items can be identified.

Second, given the developing construct of talent development, Rasch measurement provided evidence that the items selected at least formed a well-defined and hierarchically sorted set of observable characteristics that are associated with talent development as it is widely conceptualized. The implication for scoring was apparent from these results. Currently, the scoring model for the scales is to sum point scores across all of the items on a scale. This, however, does not acknowledge the different amounts of attribute (e.g., science task commitment and interest) the same category endorsement suggests across items owing to the item hierarchy. For example, on the science scale it was clear from the analysis that a category endorsement of six on *The student is curious about why things are as they are* contributed less to a quantitative manifestation of science task commitment and interest than did a category endorsement of six on *The student reads about science-related topics in his/her free time.* Such

interpretation emphasized the value of Rach measurement in validation studies as well as scoring models of rating scale instruments such as the SRBCSS-III.

Third, measurement invariance has been shown for the items on the SRBCSS-III, which indicated that higher endorsements on items was dependent on possession of a higher trait level alone for students. Closely aligned to this, specific objectivity–the idea that the difference between students was independent of the items used to compare them–was shown, as well. The overall analysis confirmed what Bond and Fox (2001) called the key question in the analysis of data: the consideration of how well a theoretical intention has been empirically realized. Indeed, the fit and reliability of the data and the validity evidence underscored such a realization.

Recommendations

The SRBCSS-III are well-functioning scales, and Rasch measurement analyses supports their use in identifying students for exposure to activities in talent development settings at varying intensities and in certain domains. However, the current study showed a few areas where development in the scales might be considered.

- An option category to indicate that a characteristic has not been observable due to the absence of opportunity will improve the functioning of some items that are showing slight misfit. Some item category options show higher-than-expected point-measure correlations, which might be caused by a forced option choice on unobservable behaviors.
- Although not large, a ceiling and floor effect was seen on some of the scales. Adding items of higher and lower difficulty should discriminate students at

those locations to improve the information that can be gained for finding the most suitable environment for growth. However, this would be most necessary in the case where the scales are used for gaining information along the range of the scales rather than in the case where they are being used to establish cut-points for talent pool identification.

- Users should be discouraged from summing item scores on a scale. Instead, support for utilizing the Rasch rating scale model to arrive at scores should be implemented with the publication. The scales are currently available from the publisher in an online environment, so analysis using Rasch measurement with meaningful scale-score reporting could be accomplished without a computational burden for users.

Additional Research

The current study was performed with a convenience sample. Although Linacre (2014a) suggests that a well-targeted group is important for researching the operationalization of a construct using Rasch measurement, a broader study using a representative sample in terms of demographics will support the findings of the current study.

The current study examined four of the six aspects of validity presented in Messick's (1989, 1995a) unified model. Additional research investigating the consequential and external aspects of validity will yield important understandings of both the SRBCSS-III and the theoretical rationale that underlie scales for identification in gifted education as talent development. The aspects of validity researched in the current

study might be grouped as those aspects of validity considered to be essential in the developmental stages of instrument construction, while studies of consequential and external aspects of validity might be grouped as relevant in what could be termed the evaluative stages of instrument construction. Related to such evaluative stage validity studies, Wolfe and Smith (2007a) specifically referred to additional aspects of validity termed responsiveness and interpretability as important to scale development. Further research of these aspects of validity should be conducted to show a) that placement in talent development programs results in intended outcomes (responsiveness validity) and b) that scores obtained through Rasch measurement are correctly interpreted by users (interpretability validity).

Additionally, similar research on the remaining scales of the SRBCSS-III will complement the work of the current study. The current study examined the psychometric properties of six scales: learning characteristics, motivation, and the four content scales. While the scales in the current study are among the most widely-used (Renzulli & Smith, 2010), the conceptualizations of giftedness in the areas of leadership, creativity, planning and other areas can be informed by a study of the remaining SRBCSS-III scales.

Conclusion

Dai and Chen (2013) noted that the fundamental problem with the conceptualization of giftedness is not a lack of definitions or attempts at identifying parameters of giftedness but instead arises from an astoundingly large number of definitions and often competing theories. Dai and Chen's commentary echoes that of Ambrose et al. (2010) who stated that a cohesive definition has yet to be found. The

current study highlighted that Rasch methods are measurement models that can provide some of the evidence needed to concretize the conceptualization of giftedness as talent development and ensure that students are identified and provided experiences consistent with theory.

Especially important in an environment where theory has yet to coalesce around a firm core of ideas, Rasch measurement illuminated the nature of the hierarchy of items on the SRBCSS-III by confirming that characteristics associated with higher ability were discriminated from those associated with lower ability. In doing so, practitioners can have confidence in locating students on the continuum of abilities to find an appropriate entry point even for students that do not yet demonstrate the behaviors at the higher end of the hierarchy. Such capability is the foundation on which talent development models build.

The property of item hierarchy speaks directly to the purpose of talent development as discussed earlier in Chapter 2. Dai and Chen (2013), for example, noted the purpose of talent development is to cultivate a range of strengths and interests–and that teachers decide the timing and trajectory. Tannenbaum (1983, 1997) suggested the role of nonintellective characteristics as well as the provision of a challenging and supportive environment, and Sternberg (1999, 2001, 2002) put forth a model that emphasizes the interactive nature of learning skills and thinking skills along with knowledge and motivation as essential to talent development toward expertise. Indeed, each of these theories highlights the contextual and changeable nature of developing talent and the importance of capitalizing on particular characteristics around which student growth can best be achieved. Again, the ability to locate students at a position

along a continuum of behaviors using the SRBCSS-III as shown in the current study supports the conceptualization of talent development described by these researchers.

Renzulli (1998) suggested that a model of giftedness or talent development must be a) researchable itself in any attempt to validate definitions used within it; and b) give direction to the programming, training, or evaluation components used to support it. Similar statements could certainly be suggested for any instrument purporting to support a model, as well. This study showed that the SRBCSS-III meets the needs of informing practitioners and researchers alike to help strengthen the foundation on which talent development models can continue to be built.

APPENDIX A: SRBCSS-III OPERATIONAL RATING SCALES USED

READING CHARACTERISTICS

The student . . .	Never	Very Rarely	Rarely	Occasionally	Frequently	Always
1. eagerly engages in reading related activities.						
2. applies previously learned literary concepts to new reading experiences.						
3. focuses on reading for an extended period of time.						
4. pursues advanced reading material.						
5. demonstrates tenacity when posed with challenging reading.						
6. shows interest in reading other types of interest-based reading materials.						
Add Column Total:						
Multiply by Weight:	1	2	3	4	5	6
Add Weighted Column Totals:	+	+	+	+	+	
Scale Total:						

MATHEMATICS CHARACTERISTICS

The student . . .	Never	Very Rarely	Rarely	Occasionally	Frequently	Always
1. is eager to solve challenging math problems. (A problem is defined as a task for which the solution is not known in advance.)						
2. organizes data and information to discover mathematical patterns.						
3. enjoys challenging math puzzles, games, and logic problems.						
4. understands new math concepts and processes more easily than other students.						
5. has creative (unusual and divergent) ways of solving math problems.						
6. displays a strong number sense (e.g., makes sense of large and small numbers, estimates easily and appropriately).						
7. frequently solves math problems abstractly, without the need for manipulatives or concrete materials.						
8. has an interest in analyzing the mathematical structure of a problem.						
9. when solving a math problem, can switch strategies easily, if appropriate or necessary.						
10. regularly uses a variety of representations to explain math concepts (written explanations, pictorial, graphic, equations, etc.).						
Add Column Total:						
Multiply by Weight:	1	2	3	4	5	6
Add Weighted Column Totals:	+	+	+	+	+	
Scale Total:						

MOTIVATION CHARACTERISTICS

The student demonstrates . . .	Never	Very Rarely	Rarely	Occasionally	Frequently	Always
1. the ability to concentrate intently on a topic for a long period of time.						
2. behavior that requires little direction from teachers.						
3. sustained interest in certain topics or problems.						
4. tenacity for finding out information on topics of interest.						
5. persistent work on tasks even when setbacks occur.						
6. a preference for situations in which he or she can take personal responsibility for the outcomes of his or her efforts.						
7. follow-through behavior when interested in a topic or problem.						
8. intense involvement in certain topics or problems.						
9. a commitment to long-term projects when interested in a topic.						
10. persistence when pursuing goals.						
11. little need for external motivation to follow through in work that is initially exciting.						
Add Column Total:						
Multiply by Weight:	1	2	3	4	5	6
Add Weighted Column Totals:	+	+	+	+	+	
Scale Total:						

Scoring:

- Add the total number of x's in each column to obtain the "Column Total."
- Multiply the "Column Total" by the "Weight" for each column to obtain the "Weighted Column Total."
- Sum the "Weighted Column Totals" across to obtain the Score for each dimension of the scale.
- Enter the Scores for each dimension on the cover sheet.

LEARNING CHARACTERISTICS

The student demonstrates . . .	Never	Very Rarely	Rarely	Occasionally	Frequently	Always
1. advanced vocabulary for his or her age or grade level.						
2. the ability to make generalizations about events, people, and things.						
3. a large storehouse of information about a specific topic.						
4. the ability to grasp underlying principles.						
5. insight into cause and effect relationships.						
6. an understanding of complicated material through analytical reasoning ability.						
7. a large storehouse of information about a variety of topics.						
8. the ability to deal with abstractions.						
9. recall of factual information.						
10. keen and insightful observations.						
11. the ability to transfer learning from one situation to another.						
Add Column Total:						
Multiply by Weight:	1	2	3	4	5	6
Add Weighted Column Totals:	+	+	+	+	+	
Scale Total:						

TECHNOLOGY CHARACTERISTICS

The student . . .	Never	Very Rarely	Rarely	Occasionally	Frequently	Always
1. demonstrates a wide range of technology skills.						
2. learns new software without formal training.						
3. spends free time developing technology skills.						
4. assists others with technology-related problems.						
5. incorporates technology in developing creative products/assignments/presentations.						
6. eagerly pursues opportunities to use technology.						
7. demonstrates more advanced technology skills than other students his or her age.						
Add Column Total:						
Multiply by Weight:	1	2	3	4	5	6
Add Weighted Column Totals:	+	+	+	+	+	
Scale Total:						

SCIENCE CHARACTERISTICS

The student . . .	Never	Very Rarely	Rarely	Occasionally	Frequently	Always
1. demonstrates curiosity about scientific processes.						
2. demonstrates creative thinking about scientific debates or issues.						
3. demonstrates enthusiasm in discussion of scientific topics.						
4. is curious about why things are as they are.						
5. reads about science-related topics in his or her free time.						
6. expresses interest in science projects or research.						
7. clearly articulates data interpretation.						
Add Column Total:						
Multiply by Weight:	1	2	3	4	5	6
Add Weighted Column Totals:		+	+	+	+	+
Scale Total:						

APPENDIX B: DESCRIPTIVE SUMMARY OF RETAINED ITEMS FROM SRBCSS-III

Scale Items	Corrected Item-Total Correlation	Mean Rating	*SD*
Reading			
R1. Eagerly engages in reading related activities	.96	4.37	1.42
R2. Applies previously learned literary concepts to new reading experiences	.96	4.35	1.29
R3. Focuses on reading for an extended period of time	.96	4.40	1.44
R4. Demonstrates tenacity when posed with challenging reading	.96	4.11	1.48
R5. Shows interest in reading other types of interest-based reading materials	.96	4.33	1.31
R6. Pursues advanced reading material independently	.96	4.19	1.49
Mathematics			
M1. Is eager to solve challenging mathematics problems (a problem is defined as a task for which the solution is not known in advance)	.87	4.10	1.34
M2. Organizes data and information to discover mathematical patterns	.90	3.84	1.38
M3. Enjoys challenging mathematics puzzles, games, and logic problems	.90	4.07	1.45
M4. Understands new mathematics concepts and processes more easily than other students	.91	3.91	1.53
M5. Has creative (unusual and divergent) ways of solving mathematics problems	.89	3.85	1.36
M6. Displays a strong number sense (e.g., makes sense of large and small numbers, estimates easily and appropriately)	.90	4.06	1.49
M7. Frequently solves mathematics problems abstractly, without the need for manipulatives or concrete materials	.88	4.05	1.45

Scale Items	Corrected Item-Total Correlation	Mean Rating	*SD*
M8. Has an interest in analyzing the mathematical structure of a problem	.88	3.52	1.43
M9. When solving a mathematics problem, can switch strategies easily, if appropriate or necessary	.90	3.94	1.42
M10. Regularly uses a variety of representations to explain mathematics concepts (written explanations, pictorial, graphic, equations, etc.)	.86	3.68	1.40
Science			
S1. Demonstrates curiosity about scientific processes	.85	4.16	1.25
S2. Demonstrates creative thinking about scientific debates or issues	.85	3.72	1.33
S3. Demonstrates enthusiasm in discussion of scientific topics	.88	4.04	1.31
S4. Is curious about why things are as they are	.81	4.21	1.28
S5. Reads about science-related topics in his/her free time	.77	3.37	1.43
S6. Expresses interest in science project or research	.84	3.84	1.43
S7. Clearly articulates data interpretation	.77	3.68	1.39
Technology			
T1. Demonstrates a wide range of technology skills	.89	3.61	1.32
T2. Learns new software without formal training	.87	3.47	1.40
T3. Spends free time developing technology skills	.85	3.30	1.32
T4. Assists others with technology related problems	.88	3.49	1.34
T5. Incorporates technology in developing creative products/assignments/presentations	.77	3.24	1.50
T6. Eagerly pursues opportunities to use technology	.87	3.72	1.38
T7. Demonstrates more advanced technology skills than other students his or her age	.87	3.36	1.42

APPENDIX C: INITIAL 73 ITEMS ON DOMAIN-AREA SCALES

Item	Item ending follows "The student…"; order on field test
Q1M	is eager to solve challenging math problems (a problem is defined as a task for which the solution method is not known in advance).
Q2R	pursues advanced reading materials independently.
Q3T	shows curiosity when new technology related equipment appears in the room
Q4S	demonstrates curiosity about scientific processes
Q5M	persists in solving math problems
Q6R	shows interest in reading other types of interest-based reading materials
Q7T	experiments with new ways to use technology
Q8S	demonstrates evidence of creative problem-solving skills
Q9M	when solving a math problem, can switch strategies easily, if appropriate or necessary
Q10R	focuses on reading for an extended period of time
Q11T	competently uses technology
Q12S	demonstrates curiosity about questions that currently have no agreed-upon answer
Q13M	has creative ways of solving math problems
Q14R	applies previously learned literary concepts to new reading experiences
Q15T	spends free time developing technology skills
Q16S	demonstrates creative thinking about scientific debates or issues
Q17M	has insightful solutions to some math problems but may not be able to fully explain reasoning
Q18R	poses original questions in reading
Q19T	transfers skills acquired from one computer software application or program to another

Item	Item ending follows "The student…"; order on field test
Q20S	demonstrates high-level questioning
Q21M	frequently solves math problems abstractly, without the need for manipulatives or concrete materials
Q22R	chooses to read during free time in class
Q23T	assists others with technology related problems
Q24S	demonstrates comfort with taking risks in science
Q25M	justifies conclusions logically and precisely.
Q26R	demonstrates tenacity when posed with challenging reading
Q27T	demonstrates more advanced technology skills than other students his or her age
Q28S	demonstrates application of scientific processes or methods to new questions or problems
Q29M	organizes data and information to discover mathematical patterns
Q30R	eagerly engages in reading related activities
Q31T	demonstrates software expertise
Q32S	demonstrates enthusiasm in discussion of scientific topics
Q33M	uses patterns to make generalizations
Q34R	enjoys and prefers reading in spare time
Q35T	experiments with new or unknown technology equipment
Q36S	clearly articulates data interpretation
Q37M	enjoys challenging math puzzles, games and logic problems
Q38T	learn new software without formal training
Q39S	approaches problems in multiple ways
Q40M	understands new math concepts and processes easier than other students
Q41T	eagerly pursues opportunities to use technology
Q42S	poses original questions
Q43M	is inquisitive about the math beyond that being studied in the classroom
Q44T	integrates information from different software programs

Item	Item ending follows "The student…"; order on field test
Q45S	collaborates well with peers in solving problems when asked to do so
Q46M	looks at the world from a mathematical perspective
Q47T	finds technology easy to use
Q48S	links science to other disciplines
Q49M	displays a strong number sense
Q50T	prefers to control the technology when working in a group
Q51S	is curious about why things are as they are
Q52M	has an interest in analyzing the mathematical structure of a problem
Q53T	demonstrates a wide range of technology skills
Q54S	designs ways to try to solve problems
Q55M	asks high-level questions such as why or what if that increase the depth and complexity of the math being studied
Q56T	assists others with software problems
Q57S	has advanced thinking skills for his/her age
Q58M	regularly uses a variety of representations to explain math concepts
Q59T	demonstrates expertise in working with technological hardware
Q60S	reads about science-related topics in his/her free time
Q61M	sees the elegant (simplest, most efficient) solution to a math problem
Q62T	integrates different technologies (e.g. video camera with computer)
Q63S	seeks opportunities to talk about one area of science of particular interest to him/her
Q64M	joins math-related activities such as Math Olympiad, math clubs, or math workshops, if the opportunity presents itself
Q65T	assists others with hardware (equipment) problems
Q66S	is tenacious without giving in to frustration
Q67M	sees the connections between different areas of math (fractions and geometry; number and algebra)
Q68T	incorporates technology in developing creative products/assignments/presentations
Q69S	enthusiastically engages in discussions of theory

Item	Item ending follows “The student…”; order on field test
Q70S	displays insight with respect to ways of addressing a problem or question
Q71S	raises thoughtful questions in class
Q72S	is an intellectual risk-taker (e.g. the student appears comfortable engaging in work that does not have one-or any!-correct answer
Q73S	expresses interest in science projects or research

APPENDIX D: POINT-MEASURE CORRELATIONS – DOMAIN SCALES RETAINED ITEMS

Scale Items	Point-Measure Correlations	
	Correlation	Expected
Reading		
R1. Eagerly engages in reading related activities	.89	.89
R2. Applies previously learned literary concepts to new reading experiences	.87	.89
R3. Focuses on reading for an extended period of time	.90	.89
R4. Demonstrates tenacity when posed with challenging reading	.90	.89
R5. Shows interest in reading other types of interest-based reading materials	.88	.89
R6. Pursues advanced reading material independently	.91	.89
Mathematics		
M1. Is eager to solve challenging mathematics problems (a problem is defined as a task for which the solution is not known in advance)	.90	.91
M2. Organizes data and information to discover mathematical patterns	.92	.91
M3. Enjoys challenging mathematics puzzles, games, and logic problems	.92	.91
M4. Understands new mathematics concepts and processes more easily than other students	.93	.91
M5. Has creative (unusual and divergent) ways of solving mathematics problems	.92	.91
M6. Displays a strong number sense (e.g., makes sense of large and small numbers, estimates easily and appropriately)	.92	.91
M7. Frequently solves mathematics problems abstractly, without the need for manipulatives or concrete materials	.91	.91

Scale Items	Point-Measure Correlations	
	Correlation	Expected
M8. Has an interest in analyzing the mathematical structure of a problem	.91	.91
M9. When solving a mathematics problem, can switch strategies easily, if appropriate or necessary	.92	.92
M10. Regularly uses a variety of representations to explain mathematics concepts (written explanations, pictorial, graphic, equations, etc.)	.89	.91
Science		
S1. Demonstrates curiosity about scientific processes	.85	.84
S2. Demonstrates creative thinking about scientific debates or issues	.87	.85
S3. Demonstrates enthusiasm in discussion of scientific topics	.89	.84
S4. Is curious about why things are as they are	.82	.83
S5. Reads about science-related topics in his/her free time	.83	.85
S6. Expresses interest in science project or research	.85	.84
S7. Clearly articulates data interpretation	.83	.85
Technology		
T1. Demonstrates a wide range of technology skills	.92	.88
T2. Learns new software without formal training	.90	.89
T3. Spends free time developing technology skills	.87	.89
T4. Assists others with technology related problems	.89	.89
T5. Incorporates technology in developing creative products/assignments/presentations	.86	.89
T6. Eagerly pursues opportunities to use technology	.89	.88
T7. Demonstrates more advanced technology skills than other students his or her age	.91	.89

REFERENCES

Ambrose, D., Van Tassel-Baska, J., Coleman, L. J., & Cross, T. L. (2010). Unified, insular, firmly policed, or fractured, porous, contested, gifted education? *Journal for the Education of the Gifted*, *33*, 453–478. doi:10.1177/016235321003300402

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park, CA: Sage Publications, Inc.

Assouline, S. G., & Lupkowski-Shoplik, A. (2012). The talent search model of gifted identification. *Journal of Psychoeducational Assessment*, *30*, 45–59. doi:10.1177/0734282911433946

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238–246.

Bloom, B. S. (1985). *Developing talent in young people*. New York, NY: Ballentine Books.

Bode, R., & Wright, B. D. (1999). Rasch measurement in higher education. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (Vol. XIV). New York, NY: Agathon Press.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: fundamental measurement in the human sciences*. Mahwah, N.J: L. Erlbaum.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: L. Erlbaum.

Borland, J. H. (2003). The death of giftedness: Gifted education without gifted children. In J. H. Borland (Ed.), *Rethinking gifted education* (pp. 97–106). New York, NY: Teachers College Press.

Borland, J. H. (2009). Myth 2: The gifted constitute 3% to 5% of the population. Moreover, giftedness equals high IQ, which is a stable measure of aptitude. *Gifted Child Quarterly*, *53*, 236–238.

Brown, S. W., Renzulli, J. S., Gubbins, E. J., Siegle, D., Zhang, W., & Chen, C.-H. (2005). Assumptions underlying the identification of gifted and talented students. *Gifted Child Quarterly*, *49*, 68–79. doi:10.1177/001698620504900107

Callahan, C., Hunsaker, S., Adams, C., Moore, S., & Bland, L. (1995). *Instruments used in the identification of gifted and talented students* (Research Monograph No. 95130). Storrs, CT: University of Connecticut.

Callingham, R., & Bond, T. (2006). Research in mathematics education and Rasch measurement. *Mathematics Education Research Journal*, *18*(2), 1–10.

Christensen, K. B., Kreiner, S., & Mesbah, M. (Eds.). (2013). *Rasch models in health*. Hoboken, NJ: John Wiley & Sons, Inc.

Cohen, L. M. (2005). Conceptual foundations for gifted education: Stock-taking. *Roeper Review*, *28*, 91–110. doi:10.1080/02783190609554344

Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Boston, MA: Pearson.

Csikszentmihalyi, M. (1997). *Creativity*. New York, NY: HarperCollins.

Dai, D. Y., & Chen, F. (2013). Three paradigms of gifted education: In search of conceptual clarity in research and practice. *Gifted Child Quarterly*, *57*, 151–168. doi:10.1177/0016986213490020

Delisle, J. (2003). To be or to do: Is a gifted child born or developed? *Roeper Review, 26*, 12–13.

Dimitrov, D. M. (2012). *Statistical methods for validation of assessment scale data in counseling and related fields*. Alexandria, VA: American Counseling Association.

Dimitrov, D. M., & Smith, R. M. (2006). Adjusted Rasch person-fit statistics. *Journal of Applied Measurement, 7*, 170–183.

Feldhusen, J. F. (2003). Beyond general giftedness: New ways to identify and educate gifted, talented and precocious youth. In J. H. Borland (Ed.), *Rethinking gifted education* (pp. 34–45). New York: Teachers College Press.

Feldman, D. H. (1988). Creativity: dreams, insights, and transformations. In R. J. Sternberg (Ed.), *The nature of creativity: Contemporary psychological perspectives* (pp. 271–297). New York, NY: Cambridge University Press.

Gagné, F. (1985). Giftedness and talent: Reexamining a reexamination of the definitions. *Gifted Child Quarterly, 29*, 103–112. doi:10.1177/001698628502900302

Gagné, F. (1995). From giftedness to talent: A developmental model and its impact on the language of the field. *Roeper Review, 18*, 103–111.

Gagné, F. (1999). My convictions about the nature of abilities, gifts, and talents. *Journal for the Education of the Gifted, 22*, 109–136.

Gagné, F. (2004). *A differentiated model of giftedness and talent (DMGT)*. Personal notes. Retrieved from http://nswagtc.org.au/images/stories/infocentre/gagne_a_differentiated_model_of_giftedness_and_talent.pdf

Gagné, F. (2009). Building gifts into talents: Detailed overview of the DMGT 2.0. In B. MacFarlane & T. Stambaugh (Eds.), *Leading change in gifted education: The festschrift of Dr. Joyce Vantassel-Baska* (pp. 61–80). Waco, TX: Prufrock Press.

Gardner, H. (1985). *Frames of mind: The theory of multiple intelligences*. New York, NY: Basic Books.

Green, K., & Frantom, C. (2002, November 14). *Survey development and validation with the Rasch model*. Presentation Manuscript, Charleston, SC. Retrieved from https://portfolio.du.edu/downloadItem/115525

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1–55. doi:10.1080/10705519909540118

Iramaneerat, C., Smith, E. V., & Smith, R. M. (2008). An introduction to Rasch measurement. In J. Osborne (Ed.), *Best practices in quantitative methods* (pp. 50–70). Thousand Oaks, CA: SAGE Publications, Inc. Retrieved from http://srmo.sagepub.com/view/best-practices-in-quantitative-methods/d6.xml

Irwin, R. J. (2007, October 20). *A psychosocial interpretation of Rasch's psychometric principle of specific objectivity*. Presented at the Proceedings of the 23rd Annual Meeting of the International Society for Psychophysics, Tokyo, Japan. Retrieved from http://www.ispsychophysics.org/fd/index.php/proceedings/article/download/284/276

Jarosewich, T. (2002). Identifying gifted students using teacher rating scales: A review of existing instruments. *Journal of Psychoeducational Assessment*, *20*, 322–336. doi:10.1177/073428290202000401

Jarosewich, T., Pfeiffer, S., & Morris, J. (2002). Identifying gifted students using teacher rating scales: A review of existing instruments. *Journal of Psychoeducational Assessment*, *20*, 322–336. doi:10.1177/073428290202000401

Jensen, A. (1997). The puzzle of nongenetic variance. In R. J. Sternberg & E. Grigorenko (Eds.), *Intelligence, heredity and environment* (pp. 42–48). Cambridge, UK: Cambridge University Press.

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards* (pp. 53–88). New York, NY: Erlbaum.

Kreiner, S., & Christensen, K. B. (2011). Item screening in graphical loglinear Rasch models. *Psychometrika*, *76*, 228–256.

Kruyen, P. M., Emons, W. H. M., & Sijtsma, K. (2012). Test length and decision quality in personnel selection: When is short too short? *International Journal of Testing*, *12*, 321–344. doi:10.1080/15305058.2011.643517

Linacre, J. M. (1997). Sample size and item calibration stability. *Rasch Measurement Transactions*, *7*, 328.

Linacre, J. M. (1998). Rasch analysis first or factor analysis first? *Rasch Measurement Transactions*, *11*, 603.

Linacre, J. M. (2000). Comparing "partial credit models" (PCM) and "rating scale models" (RSM). *Rasch Measurement Transactions*, *14*, 768.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, *3*, 85–106.

Linacre, J. M. (2004). Estimation methods for Rasch measures. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models, and applications* (p. NEED PAGES). Maple Grove, MN: JAM Press.

Linacre, J. M. (2014a). *A user's guide to WINSTEPS MINISTEP Rasch-model computer programs*. Chicago, IL: MESA Press.

Linacre, J. M. (2014b). WINSTEPS Rasch measurement (Version 3.81.0). Beaverton, OR: Winsteps.com. Retrieved from www.winsteps.com/winsteps

Lohman, D. F. (2005). An aptitude perspective on talent: Implications for identification of academically gifted minority students. *Journal for the Education of the Gifted, 28*, 333–360. doi:10.4219/jeg-2005-341

Marland, S. P. (1972). Education of the gifted and talented, volume 1: Report to the Congress of the United States by the U.S. Commissioner of Education (Vol. [Scan of Executive Summary], pp. 1–5). Washington, D.C.: U.S. Government Printing Office. Retrieved from http://www.valdosta.edu/colleges/education/psychology-and-counseling/documents/marland-report.pdf

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). New York, NY: Macmillan.

Messick, S. (1995a). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*, 5–8.

Messick, S. (1995b). Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.

Moon, T. R., & Brighton, C. M. (2008). Primary teachers' conceptions of giftedness. *Journal for the Education of the Gifted, 31*, 447–480.

National Association for Gifted Children. (n.d.). Frequently asked questions. Retrieved from http://www.nagc.org/index2.aspx?id=548

Nazim, A., & Ahmad, S. (2013). Assessing the unidimensionality, reliability, validity and fitness of influential factors of 8th grades [*sic*] student's [*sic*] mathematics achievement in Malaysia. *International Journal of Advance Research, 1*, 1–7.

Olszewski-Kubilius, P., & Lee, S.-Y. (2004). The role of participation in in-school and out-of-school activities in the talent development of gifted students. *Journal of Secondary Gifted Education, 15*, 107–123.

Olszewski-Kublius, P. (2009). The idea of "talent development." In B. MacFarlane & T. Stambaugh (Eds.), *Leading change in gifted education: The festschrift of Dr. Joyce Vantassel-Baska* (pp. 81–91). Waco, TX: Prufrock Press.

Osterlind, S. J. (2010). *Modern measurement: Theory, principles, and applications of mental appraisal* (2nd ed.). Boston, MA: Pearson Education, Inc.

Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models*. Thousand Oaks, CA: Sage Publications.

Pearse, N. (2011). Deciding on the scale granularity of response categories of Likert-type scales: The case of a 21-point scale. *The Electronic Journal of Business Research Methods*, *9*, 159–171.

Perline, R., Wright, B. D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement, 3*, 237–255.

Pfeiffer, S. (2012a). Current perspectives on the identification and assessment of gifted students. *Journal of Psychoeducational Assessment, 30*, 3–9. doi:10.1177/0734282911428192

Pfieffer, S. (Ed.). (2012b). Giftedness [Special issue]. *Journal of Psychoeducational Assessment*, *30*(1).

Pfeiffer, S. I., Petscher, Y., & Jarosewich, T. (2007). Sharpening identification tools. *Roeper Review*, *29*, 206–211.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Renzulli, J. (1978). What makes giftedness? Reexamining a definition. *Phi Delta Kappan*, *60*, 180–184.

Renzulli, J. S. (1986). The three-ring conception of giftedness: A development model for creative productivity. In R. J. Sternberg & J. Davidson (Eds.), *Conceptions of giftedness* (pp. 53–92). New York, NY: Cambridge University Press.

Renzulli, J. S. (1998). Three-ring conception of giftedness. In S. M. Baum, Reis, S. M., & L. R. Maxfield (Eds.), *Nurturing the gifts and talents of primary grade students*. Mansfield Center, CT: Creative Learning Press.

Renzulli, J. S. (2012). Reexamining the role of gifted education and talent development for the 21st century: A four-part theoretical approach. *Gifted Child Quarterly, 56,* 150–159. doi:10.1177/0016986212444901

Renzulli, J., & Smith, L. (2010). *Scales for rating the behavioral characteristics of superior students: Technical and administration manual* (3rd ed.). Waco, TX: Prufrock Press.

Renzulli, J., Smith, L., White, A., Callahan, C., Hartman, R., & Westberg, K. (2002). *Scales for rating the behavioral characteristics of superior students: Technical and administration manual*. (R. Knox, Ed.) (Rev.). Mansfield Center, CT: Creative Learning Press.

Renzulli, J., Smith, L., White, A., Callahan, C., Hartman, R., Westberg, K., … Sytsma Reed, R. (2013). *Scales for rating the behavioral characteristics of superior students (Renzulli scales)*. Waco, TX: Prufrock Press.

Renzulli, J. S., Siegle, D., Reis, S. M., Gavin, M. K., & Systma Reed, R. E. (2009). An investigation of the reliability and factor structure of four new scales for rating the behavioral characteristics of superior students. *Journal of Advanced Academics, 21*, 84–108.

Robinson, A., Shore, B. M., & Enersen, D. L. (2007). *Best practices in gifted education: an evidence-based guide*. Waco, TX: Prufrock Press.

Robinson, N. M. (2009). Traditional and alternative assessment of intellectual and academic aptitude. In B. MacFarlane & T. Stambaugh (Eds.), *Leading change in gifted education: the festschrift of Dr. Joyce Vantassel-Baska* (pp. 243–252). Waco, TX: Prufrock Press.

Schroth, S. T., & Helfer, J. A. (2009). Practitioners' conceptions of academic talent and giftedness. *Journal of Advanced Academics*, *20*, 384–403.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*, 107–120. doi:10.1007/s11336-008-9101-0

Smith, R. M. (2003). *Rasch measurement models: Interpreting WINSTEPS/BIGSTEPS and FACETS output*. Maple Grove, MN: JAM Press.

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. New York, NY: Cambridge University Press.

Sternberg, R. J. (1995). *A triarchic approach to giftedness* (Research Monograph No. 95126). Storrs, CT: University of Connecticut.

Sternberg, R. J. (1999). Intelligence as developing expertise. *Contemporary Educational Psychology*, *24*, 359–375.

Sternberg, R. J. (2001). Giftedness as developing expertise: A theory of the interface between high abilities and achieved excellence. *High Ability Studies*, *12*, 159–179. doi:10.1080/13598130120084311

Sternberg, R. J. (2002). Giftedness according to the theory of successful intelligence. In N. Colangelo & G. A. Davis (Eds.), *Handbook of gifted education* (3rd ed., pp. 88–99). Boston, MA: Allyn & Bacon.

Subotnik, R. F. (2003). A developmental view of giftedness: From being to doing. *Roeper Review, 26*, 14–15. doi:10.1080/02783190309554233

Subotnik, R. F., Olszewski-Kubilius, P., & Worrell, F. C. (2011). Rethinking giftedness and gifted education: A proposed direction forward based on psychological science. *Psychological Science in the Public Interest*, *12*, 3–54. doi:10.1177/1529100611418056

Tannenbaum, A. (1983). *Gifted children: Psychological and educational perspectives*. New York, NY: Macmillan.

Tannenbaum, A. (1997). The meaning and making of giftedness. In N. Colangelo & G. A. Davis (Eds.), *Handbook of gifted education* (2nd ed., pp. 27–41). Boston, MA: Allyn & Bacon.

Tannenbaum, A. (2003). Nature and nurture of giftedness. In N. Colangelo & G. A. Davis (Eds.), *Handbook of gifted education* (3rd ed., pp. 45–59). Boston: Allyn and Bacon.

Uebersax, J. (2006). *Likert scales: Dispelling the confusion*. Retrieved from http://john-uebersax.com/stat/likert.htm

U.S. Department of Education. (2014, March 6). Jacob K. Javits gifted and talented students education program. Retrieved from http://www2.ed.gov/programs/javits/funding.html

VanTassel-Baska, J. (Ed.). (2008). *Alternative assessments with gifted and talented students*. Waco, TX: Prufrock Press.

Witty, P. A. (1958). Who are the gifted? In N. B. Henry (Ed.), *Education of the gifted. 57th yearbook of the national society for the study of education: Part 2* (pp. 41–63). Chicago, IL: University of Chicago Press.

Wolfe, E. W., & Smith, E. V. (2007a). Instrument development tools and activities for measure validation using Rasch models: Part I - instrument development tools. *Journal of Applied Measurement*, *8*, 97–123.

Wolfe, E. W., & Smith, E. V. (2007b). Instrument development tools and activities for measure validation using Rasch models: Part II - validation activities. *Journal of Applied Measurement*, *8*, 204–234.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, *14*, 97–116.

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

BIOGRAPHY

David Alan Nelson received his Bachelor of Education from The University of Toledo in Toledo, Ohio in 2000. He received his Master of Science from The University of Maryland in College Park in 2009. He has worked as a science teacher and a teacher specialist in assessment in Calvert County Public Schools in southern Maryland since 2001.