VALIDATING A MATHEMATICS INTERIM ASSESSMENT WITH COGNITIVELY DIAGNOSTIC ERROR CATEGORIES

A Dissertation

Presented to the Faculty of the Curry School of Education

University of Virginia

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By

Christine C. Hutchison, B.A., Lynchburg College

May, 2014

Abstract

J. Patrick Meyer, III

The stressors of the No Child Left Behind Act have thrust educators into a data-driven accountability culture. As school divisions are racing to keep up with increasingly higher achievement demands, educators are scrambling to find testing and instructional methods for improving mathematics achievement *prior* to students sitting for end-of-grade (EOG) and end-of-course (EOC) tests. Over the past several years, interim assessments have emerged as a possible solution, although there is a paucity of empirical research to support interim assessments as vehicles for improving mathematics achievement. The purpose of this mixed methods study was to create and validate a 7th grade mathematics interim assessment which incorporated cognitively diagnostic error categories. The interim assessment followed an ordered multiple-choice test design where distractors represented students' common errors. Inspiration for the development of the error categories came from the cognitive and school improvement literature. The error categories comprise: conceptual, procedural, and attention errors. Validity evidence was gathered from qualitative sources (i.e., student cognitive think-alouds, expert teacher reviews), and quantitative sources (i.e., classical test theory analysis, distractor analysis, differential item functioning, and a partial credit item response theory analysis). Results suggest that there is validity evidence to support the development of the cognitively diagnostic error categories and the overall test design. Of the three error categories, the attention error category was the most problematic and erratic. Validity evidence to

support the ordering of the error categories was not consistent. More research needs to be done in the development of the attention error category and the ordering of all three error categories. Limitations to the study and opportunities for future research were discussed.

Department of Educational Leadership, Foundations, and Policy
Curry School of Education
University of Virginia
Charlottesville, Virginia

APPROVAL OF THE DISSERTATION

This dissertation, Validating a Mathematics Interim Assessment with Cognitively Diagnostic Error Categories, has been approved by the Graduate Faculty of the Curry School of Education in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

_____
J. Patrick Meyer, III

_____
Timothy R. Konold

_____
Robert Q. Berry, III

_____
Robert McNergney

_____
Date of Defense

Dedication

I dedicate this dissertation to my loving husband Jim who has supported me over the many long years it took for this dissertation to come to fruition. I know that waiting for me to finish this degree has been difficult on many levels, but I greatly appreciate your love, devotion, and support. Thank you!

Acknowledgments

My husband, Jim has told me on many occasions that a dissertation is not accomplished by one person. It requires the support of so many others. Jim is so right.

With this thought in mind, I want to thank my husband, Jim whom I adore and my children Sean, Paul, and Rachel for their unrelenting love and support over these past several years. You have never stopped encouraging me. Thanks so much! I could not have done this without you.

Second, I would like to thank Patrick Meyer, my doctoral advisor who has been there for me since my doctoral studies began. Before I took "Measurement Theory," you told me that I would love the class. You were so right. Being your student was a joy. There's no doubt that your influence brought me to where I am today. Thank you!

Hearty thanks also go to my doctoral committee for your invaluable input into the dissertation process and my doctoral studies. I could not have had a better committee. Thanks Tim Konold for your influence throughout my doctoral studies. You are a wealth of information in educational research and statistics. And you made upper-level statistics so approachable! Bob McNergney, I so enjoyed all of our conversations about life and, of course, education. You are a delight! Robert Berry, I am so happy that you agreed to be on my doctoral committee. Your feedback has been welcomed and appreciated. Thank you!

Several people also helped me with various aspects of the dissertation process: Thanks go to Matt Carroll for providing the raw data in a format that I could use. Thanks Tim Luken and Jennifer Elliott for transcribing my interviews. What a relief it was to

know that someone else could take care of this task so that I could focus on the data analysis and writing. You both were such a blessing. Thanks, too, to all of my friends and extended family who are too numerous to mention. Know that your love and support do not go unnoticed. Finally, thanks to all of the struggling mathematics students that I have taught that provided the initial inspiration for this research. You are the reason that this research needs to be done.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter I: Validating a Mathematics Interim Assessment with Cognitively Diagnostic Error Categories

With the enactment of the No Child Left Behind Act of 2001(No Child Left Behind Act of 2001, 2002) all states are required to establish accountability systems with the express purpose of closing the achievement gap among all students, especially marginalized subgroups. Consequently, all states administer high-stakes summative assessments which reflect state content standards in reading and mathematics. In the elementary and middle school, these assessments are referred to as end-of-grade (EOG) assessments while in the high school setting these assessments are called end-of-course (EOC) tests.

In the eleven years since NCLB has become law, education has been transformed into a data-driven culture. Educators are expected to use EOG and EOC summative assessments as barometers of student achievement as well as indicators of how achievement and instruction might be improved. But, educators and policymakers know that EOC/EOG assessments do not "provide instructionally useful information" (Perie, Marion, & Gong, 2009, p. 5). In fact, many schools have not been able to maintain the increasingly high achievement benchmarks required to attain adequately yearly progress (AYP) and school accreditation. Without this consistent attainment of AYP, some schools have been forced into school improvement by their State Departments of Education. If school improvement has not ameliorated the school's achievement, the school becomes subject to sanctions and possible government takeover (Hu, 2011; "School Boards File Lawsuit School Takeover is Unconstitutional," 2013). School takeovers have occurred in many states: New York, California, Connecticut, New Jersey, and Pennsylvania (Hu,

2011; Rundquist, 2013). Not surprisingly, this legislation has caused considerable angst among teachers and administrators alike.

In response to the stressors of the NCLB accountability climate, President Obama has called for a reauthorization of the Elementary and Secondary Education Act (ESEA) of 1965 (United States Department of Education [US Department of Education], 2010). Obama's plan eliminates AYP benchmarks with more flexible school reform measures. His plan addresses several key priorities: a focus on all students being college and career ready through new courses, tests, and teacher professional development centered on new college and career ready standards; a revision of teacher and principal evaluation measures based on teacher observations and student growth models; greater support for charter schools and other innovative instructional plans which address continuous school improvement and closing the achievement gap (US Department of Education, 2010, pp. 3-6). Clearly, accountability is not going away. Despite President Obama's push for reauthorization of the ESEA, educators still require instructionally meaningful methods to measure and increase student achievement.

Given NCLB and the reauthorization of the ESEA, some school districts have implemented interim assessments to provide teachers with instructionally useful data about students' achievement *prior* to their sitting for the state EOG/EOC assessment. Interim assessments are defined as assessments that lie characteristically *between* formative and summative assessments (Perie et al., 2009; Perie, Marion, Gong, & Wurtzel, 2007). Unlike formative assessments, interim assessments are not classroom assessments (Clune & White, 2008; Perie et al., 2009; Perie, Marion, Gong, & Wurtzel, 2007). They are standardized assessments developed by testing companies, school

district content specialist(s), or a combination of teachers and district level content specialist(s). Although teachers may participate in the development of interim assessments, they are typically not the sole authors. Whereas formative assessments are administered during or immediately following a specific curricular unit, interim assessments are typically administered one to three times a year. As such, feedback from interim assessments is less immediate to instruction. Because interim assessments are administered less frequently than classroom assessments, they can examine student retention of key content and student growth as teachers prepare students for the EOC/EOG assessment.

Background

Little empirical research has occurred with interim assessments. Much of the research has relied on teacher observations, interviews, and surveys (Christman et al., 2009; Clune & White, 2008; Goertz, Oláh, & Riggan, 2009; Marsh, Pane, & Hamilton, 2006). Several of these studies revealed that teachers alter their instruction in response to interim assessment data (Christman et al., 2009; Clune & White, 2008), although there is substantial variability in how effective teachers are in their data analysis and interpretation (Goertz et al., 2009). Fewer studies have investigated the effect of interim assessments on student achievement. The Carlson et al. (2011) study is one of the first large-scale empirical studies which suggests that interim assessments are a viable means of improving student achievement in mathematics. The Carlson interim assessments mirrored the state test blueprint and were administered as quarterly, predictive assessments. Training was provided to teachers and administration in data analysis, data interpretation, and the data-driven reform process. Despite the significant contributions

of the aforementioned studies, no research has explored what interim assessment framework may be cognitively and instructionally meaningful for teachers. Cognitive diagnostic assessments offer a possible interim assessment framework to support these goals.

Cognitive diagnostic assessments (CDAs) are assessments of student learning which diagnose student "knowledge structures and cognitive processing skills" so that remediation is informed (Leighton & Gierl, 2007b, p. 3; Nichols, 1994). Leighton and Gierl (2007a) suggest that if a cognitive model is not empirically derived, then it cannot support diagnostic inferences. Nichols (1994) submits that the underlying cognitive theory of CDA is used to *generate* assessments and predict results. Assuming that a CDA meets both conditions set forth by Leighton, Gierl, and Nichols, the resulting data should allow teachers to "alter student misconceptions and faulty strategies" (Leighton & Gierl, 2007b, p. 6). However, despite these theoretical definitions, no clear testing framework for CDA achievement tests has been established.

Several researchers have created CDA models as a paradigm for diagnosing student content strengths and weaknesses and to potentially inform instruction (Briggs, Alonzo, Schwab, & Wilson, 2006; Embretson, 1998; Gierl, Leighton, & Hunka, 2007; Mislevy, Almond, & Lukas, 2003; Mislevy & Haertel, 2006; Rupp & Templin, 2008; Rupp, Templin, & Henson, 2010; Tatsuoka, 2009). The challenge is to find a testing framework that is not too fine or large grain so that the test data are not overwhelming to teachers while simultaneously being rich in diagnostic data. Six approaches to CDAs are discussed in this study: the Rule-Spaced Model (RSM), the Attribute Hierarchy Model

(AHM), the Diagnostic Classification Model (DCM), Evidence-Centered Design (ECD), the Cognitive Design System (CDS), and Ordered Multiple Choice (OMC) assessments.

The RSM is a statistical model that combines error theory and item response theory (IRT) to classify and diagnose student's cognitive errors (Tatsuoka, 1983, 1986, 2009). The RSM is comprised of two stages, the selection of feature variables and statistical pattern classification. The selection of feature variables is supported by Q matrix theory and the development of several supporting matrices. Oftentimes the RSM is retrofitted to an existing test which compromises the cognitive diagnostic capability of the assessment (Gierl, 2007; Gierl et al., 2007). Moreover, the RSM offers the educator a fine-grained diagnosis of student content strengths and weaknesses which would likely not be practical for teachers.

Like the RSM, the AHM employs Q matrix theory and attributes which are developed by content experts (Gierl, 2007; Gierl et al., 2007; Leighton, Gierl, & Hunka, 2004). AHM attributes are represented in the same type of matrices seen in Tatsuoka's RSM. The primary differences between the RSM and the AHM are (a) the AHM is applied *a priori* to a test and (b) the AHM focuses on student content mastery opposed to student error analysis. An emphasis on content strengths may leave gaps in teachers' understanding of student weaknesses as content weaknesses are not necessarily an absence of content strengths. Students have content weaknesses for a variety of reasons and the AHM does not appear to help in illuminating what those weaknesses are. Furthermore, the AHM is likely too mathematically sophisticated for most teachers to comprehend. The fine-grain diagnosis of students' achievement using the AHM would

likely overwhelm teachers. Finally, the AHM is not recommended for achievement tests (Gierl et al., 2007).

The next model, the DCM is a statistical model that *predicts* student performance according to a set of mastered attributes (Rupp & Templin, 2008; Rupp et al., 2010). Like AHM, the DCM employs a Q matrix to specify the item attribute relationship where attributes are assigned a priori. However, because the DCM allows any given item to have more than one latent skill load on it, the items are multidimensional. For instance, if one attribute is addition and another is subtraction, one item might assess both the addition and subtraction attributes. Unlike most assessment models, the DCM does not provide a scaled score, but rather a profile of mastered skills or attributes. The student mastery profiles are presented as *probabilities* of mastery. Despite the attribute mastery profiles, DCMs focus on why a student is not performing well. Because this model is centered on skills acquired rather than a cognitive processing diagnosis, it is not an appropriate model for this study.

Following is ECD which is a construct-centered approach that focuses on the accumulation of evidence to support student inferences. Thus, ECD addresses the validity of a test's scores. ECD divides test development into four models: student model, evidence model, task model, and assembly model (Gorin, 2007; Mislevy et al., 2003). The student model characterizes a student's mastery of specific skills linked to the test's purpose. The evidence model describes the observable behaviors required to support the student model while the task model defines the kind of task required to elicit item mastery. Finally, the assembly model expresses how the student, evidence, and task models work together to present the final assessment. Although the ECD provides a

comprehensive approach to cognitive test development and test validity, it also appears to be a renaming of traditional test design principles. Hence, the ECD does not appear to offer any new information for this study.

The fifth model, the CDS merges cognitive principles with test design in a fashion that may not be overwhelming to teachers. With CDS, cognitive theory *precedes* test design and item development (Embretson, 1998, 1999, 2010; Embretson & Gorin, 2001; Gorin, 2007). Psychometric models are used to evaluate the model fit of the item responses. Items with good model fit provide strong evidence of the construct being measured. On the other hand, items with poor model fit need to be reviewed for item refinement or the construct needs to be reconsidered. CDS is typically used for ability measures, such as spatial reasoning tasks. Some of Embretson's ability tests are multiple-choice assessments. Embretson's most prominent finding from her multiple-choice assessments was that the decision process was impacted by the nature of the distractors (Embretson & Gorin, 2001, p. 360).

The last model is OMC assessments, which given their multiple-choice format already is familiar to teachers. OMC assessment researchers recommend that distractors should be written using students' common errors and misconceptions (Briggs et al., 2006); however, this model does not include students' thinking processes in the composition of test items. In addition, the distractors do not represent error categories, but rather *levels* of student understanding. Although the OMC assessments are the closest test framework to the interests of this study, changes still need to be made so that they are suitable as cognitively diagnostic interim assessments.

Although some empirical studies have investigated these assessment types separately, I have found no empirical studies that have validated an interim assessment that is cognitively diagnostic. This mixed methods study merges interim assessments, CDA models, cognitive psychology, and mathematics cognition using a framework built around the following principles. First, teachers are demanding that any required assessments have "maximum instructional value" (Huff & Goodman, 2007, p. 24). If interim assessments could point specifically to where student misconceptions lie and where deficits are in students' cognitive processes, interim assessments could foreseeably render "maximum instructional value." Second, interim assessments must incorporate student misconceptions in the item scoring. Third, most interim assessments are multiple-choice tests where item responses are scored dichotomously. The resulting scores describe student's correct responses, but reveal nothing about students' incorrect responses. As such, these assessments lose valuable data about student misconceptions making instructional modification difficult (Black & Wiliam, 1998) and fueling criticisms for why they cannot be used for cognitive diagnosis (Hermann-Abell & DeBoer, 2011). Fourth, some CDA models are fine-grain tests that potentially inundate teachers with too much data. For instance, if a teacher administers a 25-item mathematics interim assessment to her 120 secondary students and each item has 4 response options, she has 25 data points per student. This translates into 3,000 data points per interim assessment. Clearly, the volume of this data is overwhelming to teachers who are attempting to remediate student cognitive weaknesses and extend student understandings. My goal is to develop a mathematics interim assessment framework that integrates each of these principles.

Inspiration for how to develop a mathematics interim assessment to meet each of these principles comes from the cognitive and interim assessment literature (Baddeley, 2007; Bjorklund, 2005; Dehn, 2008; Feifer & De Fina, 2005; Goertz, Oláh, & Riggan, 2009; Matlin, 2002; Mazzocco & Devlin, 2008). For instance, Goertz et al. (2009) investigated how teachers used multiple-choice mathematics interim assessments to modify instruction. These teachers developed four error categories to explain student performance: a *procedural-conceptual continuum*, *conceptual understanding*, *other cognitive weaknesses* which included test anxiety, difficulty maintaining attention, and weak reading ability, and *contextual diagnoses* which were outside the realm of school influence. Several components of these error categories fit with an information processing approach to cognition: *procedural-conceptual continuum, conceptual understanding*, and *attention* (Baddeley, 2007; Bjorklund, 2005; Dehn, 2008; Feifer & De Fina, 2005; Matlin, 2002).

To capture the spirit of the aforementioned categories and provide a mechanism for teachers to more easily differentiate and remediate instruction, I redefined the error categories as *procedural knowledge, conceptual or declarative knowledge,* and *attention*. A brief definition of each error follows.

Attention errors suggest lapses in selective and/or sustained attention abilities (Baddeley, 2007; Bjorklund, 2005; Feifer and De Fina, 2005; Matlin, 2002; Sergeant, 1996). I submit that attention errors are more about what students did not do in their problem solving, rather in what they did do. In other words, attention errors are sins of omission instead of commission. For instance, they may have *left off* the last step when solving a problem, thinking they had completed the entire series of steps. This *leaving*

off of the last step may be due to a lapse in sustaining their attention on the mathematics problem solving. In this case, the formula(s) chosen are correct, and the procedures and calculations are correct.

Procedural errors are related to procedural knowledge and are defined as calculation errors or missteps in problem solving. For example, in the calculation of a multi-step problem, perhaps the student begins the problem correctly and then makes a calculation error further in their problem solving. Or, in a word problem, maybe they understood what the problem was asking, selected the correct formula, but then they made mistakes using the formula. Or, perhaps the student calculated the slope of a line as $\frac{\Delta x}{\Delta y}$, rather than $\frac{\Delta y}{\Delta x}$.

Conceptual/declarative errors are tied to conceptual/declarative knowledge, which refers to "knowledge about facts and things" (Matlin, 2002, p. 254). In a mathematics context, conceptual errors can be defined as mathematics vocabulary, mathematics facts, mathematics rules, mathematics notation and their meanings, and the selection of an appropriate formula or operation for a given problem. Many times the errors refer to the student's absence of some factual information. For example, suppose a student is asked to graph a series of points in a Cartesian coordinate plane. If the student reverses the x- and y-coordinates for *all* of the points, it could be deduced that the student does not know which axis is the x-axis or y-axis. On the other hand, if the student did not reverse all of the coordinates, it is possible that this represents a procedural error. Conceptual errors also encompass conceptual understanding, but some resources are not clear on exactly what this entails.

These new error categories can potentially link students' thinking processes with their acquired skills. As a result, the error categories can give direction to teachers in how they remediate student content weaknesses. For example, if a student has more conceptual errors, a teacher might use manipulatives or other hands-on instructional strategies to remediate their conceptual misunderstandings. Alternatively, if a student has more procedural errors, a teacher might focus on using foldables, flow charts, or thinking maps to review the correct process for solving a specific type of problem. On the other hand, if the student has made more attention errors, the teacher might work on helping the student with planning and executive function skills so that the student can be more successful in his or her mathematical problem solving.

To align the error categories with item scoring, it would be helpful to order these categories according to their degree of correctness. Several researchers' work gives direction to how the ordering of these error categories could be conceptualized.

Mazzocco and Devlin's (2008) research compared students with low mathematics achievement (LA) to those with mathematical learning disabilities (MLD). Their research showed that in general students with MLD had a "weak rational number sense and inaccurate beliefs about rational numbers" whereas students with LA exhibited a partial understanding of fractions and decimals with a propensity to memorize labels, procedures, and fraction to decimal equivalencies without a clear understanding of fundamental concepts (Mazzocco & Devlin, 2008, p. 690). This study not only gives credibility to the creation of conceptual and procedural error categories, but it also suggests that students with less mathematical skill (MLD) make more conceptual errors than students with more mathematical skill while students with slightly more skill (LA)

have some conceptual and procedural understanding. Other research studies are consistent with this generalization (Geary, Hoard, & Bailey, 2011; Mazzocco, Myers, Lewis, Hanich, & Murphy, 2013).

Baddeley's (2007) model of working memory provides further inspiration for an attention error category. His model is divided into three components: the phonological loop, the visuo-spatial sketchpad, and the central executive. Depending upon the nature of a given mathematics problem, each of these working memory components could be involved in mathematical problem solving. However, attentional capacities reside within the central executive component and are the most "crucial feature of working memory" (Baddeley, 2007, p. 124). The central executive is "crucial" because it determines which information in a mathematics problem should receive attention and which should not (Feifer & De Fina, 2005). Furthermore, Baddeley (2007) and Feifer and De Fina (2005) argue that not only is the central executive critical in directing, shifting, and sustaining attention, but it is also important in the inhibition of negative distractors and the selection of necessary strategies to execute a cognitive task. The central executive, therefore, orchestrates the action required in working memory and, in turn, mathematical problem solving. Since attention is more about focus and orchestration and less about the storage of memory components, perhaps attention errors are lesser errors than procedural and conceptual errors because they are not about stored memory components.

Based on the aforementioned research, I am positing that attention errors are the least serious of the three error categories, followed by procedural errors, and then conceptual errors. Once the data is collected, the exact ordering of the error categories will be validated.

The purpose of this study is to create and validate an interim assessment for 7th grade mathematics. The assessment will consist of ordered multiple-choice categories with distracters that contain common student misconceptions. The error categories will be linked to cognitive processes and will comprise: attention, procedural knowledge, and declarative or conceptual knowledge. Through error analysis, educators will be able to determine student's error patterns so that teachers will be better equipped to remediate their misconceptions and faulty strategies and extend their understandings.

This study will answer the following research questions:

(1) What validity evidence from the expert reviews and cognitive interviews supports the error categories?

(2) What is the relationship between students' problem-solving errors and teachers' perceptions of students' problem-solving errors?

(3) What is the item response theory evidence to support the OMC interim assessment framework?

(4) Do the errors made by advanced 6th grade mathematics students differ from those made by general 7th grade mathematics[1] students?

(5) Do the errors made by special education mathematics students differ from those made by regular mathematics students?

---

[1] In this school division many SOL courses are divided into two sections, "advanced" and "general." In this division, 7th grade mathematics SOLs are taught in two courses, advanced 6th grade mathematics and general 7th grade mathematics. Students enrolled in advanced 6th grade mathematics are presumed to have higher ability levels while those in general 7th grade mathematics are presumed to have lower ability levels. Thus, this terminology is division specific.

Methods

Because the purpose of this study is to create and validate an interim assessment for 7th grade mathematics and to use the resulting test scores to inform instruction, validity evidence was required from multiple sources to support inferences drawn about this population (Crocker & Algina, 2008; Haladyna, 2004; Haladyna & Rodriguez, 2013; Kane, 2009). As suggested by Haladyna (2004), much of this validity evidence came from a study of the item development procedures and an item analysis. Additional evidence was gathered from expert teachers' reviews and student interviews. Therefore, a mixed-methods research design was employed because quantitative and qualitative methods were necessary to answer the research questions. Analyses were performed separately and then *mixed* during the discussion and interpretation of the data. Ultimately, the validation process was a joining together of the test's purpose, its inferences about the population, and the evidence gathered to support those inferences (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Kane, 2006; Schmeiser & Welch, 2006).

Test development. One middle school mathematics teacher and I created the initial interim assessment test items. Each test writer used the Virginia SOL Test Blueprint (Virginia Department of Education [VDOE], 2009a), 7th grade Mathematics Curriculum Framework (Virginia Department of Education [VDOE], 2009b), and the local 7th grade mathematics pacing guide as guides for test item content development. In addition, the test writers followed the item-writing guidelines promulgated throughout the

theoretical and empirical research literature (Haladyna, 2004; Haladyna & Downing, 1989a; Haladyna & Downing, 1989b).

We created a 25-item interim assessment with ordered multiple-choice (OMC) error categories such that distracters represented common student misconceptions. This interim assessment represented the third 9 weeks interim test for a Virginia school division. The interim assessment addressed SOLs: theoretical and experimental probability (7.9), compound probability (7.10), statistics (7.11), relations (7.12), writing and evaluating algebraic expressions (7.13), properties of operations (7.16), arithmetic and geometric sequences (7.2), and solving linear equations (7.14). Seventeen test items were devoted to these SOLs out of the total of 25 test items. The remaining 8 items constituted common items from the previous interim assessment. After the second 9 weeks interim test administration, a frequency analysis was used to determine the lowest performing items for the division. The frequency analysis measured the number and percent of students who correctly responded to each item. The 8 items with the lowest percent correct formed the basis for the common items on the subsequent interim assessment. Common items did not refer to the same test items, but rather *similar* test items. Common items were defined as (a) items which used the same question stem with different numbers, (b) items in which the sequence of the response options changed so that if option A was previously correct, this was no longer the case, and (c) items which retained the same location within the overall test. I am hypothesizing that this common item model was a sufficient measure of student remediation and growth in achievement. The common item SOLs included: proportional reasoning (7.4), volume and surface area

of rectangular prisms and cylinders (7.5), similar quadrilaterals and triangles (7.6), properties of quadrilaterals (7.7), and transformations of polygons (7.8).

Once the first draft of each test was developed, middle school mathematics teachers from the division reviewed the items and key through a peer review process to ensure that the items were aligned with the 7th grade mathematics SOLs and the division's pacing guide. Teachers were asked to confirm that the key was accurate and that all test items were clear, free from errors, and appropriate for 7th grade students. No revisions to items or the key were required before test implementation.

Each item included three distractors which encompassed students' common mistakes or misconceptions. Based on the cognitive research previously discussed, I propose that the error categories attention, procedural knowledge, and conceptual knowledge were *sufficient* in describing students' common mathematics mistakes. Although these three error categories might be satisfactory, I do not anticipate that they describe all nuances of mathematics cognition. Other categories may emerge in this study, but at this point, these three error categories appear to be the most salient in describing students' mathematics problem solving and the source of their common errors.

Errors were ordered according to their degree of correctness. Students' scores for each item were represented as partial credit scores. Correct responses were scored as 3, attention errors as 2, procedural errors as 1, and conceptual/declarative errors as 0.

These error categories are not believed to be mutually exclusive (Kruschke, 2005; Rittle-Johnson & Siegler, 1998). The overlap and influence of one error category on another could make the assignment of error categories arduous. Thus, the assigned error category represented the *primary* or overarching cognitive error. In this way, the most

prominent cognitive feature of the error should be recognized. Correct error category assignment is very important to the eventual success of teachers' remediation efforts. Incorrect error category assignment could result in a teachers' misunderstanding of student mistakes and the selection of inappropriate remediation interventions.

Cognitive interviews and expert teacher reviews. To validate the development of the error categories, 12 students and 3 teachers participated in retrospective cognitive think-alouds (TA). Student TAs centered on students re-enacting their mathematics problem solving. Probes constituted questions such as, *"tell me why you moved from this step to that step?"* or *"can you show me what you were thinking?"* or *"why did you not select answer A?"* Answers to some of these probes could illuminate a student's problem-solving rationale and perhaps pinpoint why students struggle mathematically. The resulting data was used to confirm the development and assignment of the error categories. Teacher TAs focused on their perceptions of student errors. The teacher TA data allowed for the comparison of teachers' pre-conceptions of student errors with students' explanations of their reasoning.

Three different middle school mathematics teachers participated in the rating of the error categories. The teacher raters were trained according to the error category rubric. The training constituted a theoretical explanation of each error category and its relationship to cognitive processing. Several test examples were provided of incorrect responses and their error category assignment. The teacher raters were given the opportunity to practice categorizing errors on several items before the rating data was collected. The teachers rated all 25 test items using the interim assessment items and the

answer key *without* error categories. Since each item had three incorrect responses, each teacher rated 75 incorrect response options.

Participants. All advanced 6th grade and general 7th grade mathematics students were administered the interim assessment. However, nine advanced 6th grade and general 7th grade mathematics students participated in the student TAs. Three mathematics teachers participated in the teacher TAs and three different mathematics teachers were teacher raters. Each teacher participant had a minimum of three years of teaching experience. As such, teacher participants had an in-depth understanding of common student misconceptions and middle school mathematics.

Data collection procedures. Parental consent forms were distributed to all advanced 6th grade and general 7th grade mathematics students at three middle schools prior to interim assessment administration. A stratified random sample of students was selected from those forms in which parents bestowed consent. Stratification occurred in two ways: middle school matriculation and course enrollment (i.e, advanced 6th grade mathematics and general 7th grade mathematics). This double stratification ensured that students from each course were represented in a given school's random sample. Because three students were selected from each school, a total of nine students were selected for the TAs. Student assent procedures (i.e., written and verbal assent) were followed to ensure the willingness of each student to participate in the cognitive interview process.

The sampled TA students were interviewed in a secure testing location on the school campus. Student 1 at each school was interviewed for items 1-8, student 2 at each school was interviewed for items 9-16, and student 3 was interviewed for items 17-25. A student think-aloud protocol was utilized giving each interview a consistent structure.

Each think-aloud was audio-taped and transcribed to ensure the interview methodology was followed. The think-aloud interview was untimed allowing students ample time to problem solve.

Subsequently, three teacher TAs were administered in the teacher's classrooms during their planning periods. The teacher TAs were audio-taped and transcribed to ensure the interview methodology was followed. Educators examined the same test items the sampled students were given.

Finally, the teacher raters were trained using the training protocols. Once the teacher raters were comfortable with how to rate the distractors, they each rated all 25 test items. Teacher raters were permitted to use their training resources to better assist them in assigning error categories to each distractor.

Student group comparisons according to types of errors. In the current accountability climate, teachers are required to use test data to drive instruction, but many teachers are in a quandary how to achieve this expectation (Mandinach & Honey, 2008; Marsh, et al., 2006; Young, 2006). Anecdotally, these teachers may make comparisons between some of their student subgroups (e.g., special education students, second-language learners, disadvantaged students) wondering *what are the cognitive strengths of some of these groups?, what are their shared cognitive weaknesses? should I use different instructional strategies for one group versus another to increase achievement?* If, for instance, teachers had evidence that special education students make more conceptual errors as a group than general education students[2], this data could be helpful in teachers'

[2] General education is a division-specific term that signifies those students who do not receive special education or 504 accommodations.

instructional planning and remediation. It is questions like these that suggest an investigation needs to be made comparing the student groups in this study.

The student participants in this study were divided into several student groups: advanced 6th grade versus general 7th grade mathematics students and general education versus special education mathematics students. I was interested in discerning if either of these student groups differs in the types of errors made. If so, this data would provide teachers with valuable lesson planning/remediation data. For example, if general 7th grade mathematics students as a group made more conceptual errors than advanced 6th grade mathematics students, this data would suggest that teachers place greater emphasis on conceptual knowledge constructs (e.g., mathematics vocabulary, mathematics facts, mathematics notation, and general mathematics concepts).

Data analysis. An item analysis was performed using the psychometric program jMetrik (Meyer, 2002). The item analysis comprised a classical test theory analysis, distractor analysis, differential function analysis (DIF), and item response theory (IRT) analysis of polytomous items. Since the test items are polytomous where "partial credit [is awarded] for partial success," the IRT partial credit model was employed (Masters, 1982, p. 150). An important contribution of an IRT analysis is the item map (Wilson, 2005). Here the item map served two functions: it helped validate the ordering of the error categories and it provided teachers an additional diagnostic tool of students' content strengths and weaknesses. Furthermore, the error categories were also validated using a chi-square analysis, which compared the IRT theta values to the error category assignments on the test key.

Transcriptions of the student and teacher TAs were analyzed for patterns and trends using case-ordered matrices (Miles & Huberman, 1994). Student and teacher responses were analyzed to determine how well they fit with the error category descriptions. Descriptive statistics were calculated to characterize the quality of the student mathematics talk and teacher's expert review of students' problem-solving behaviors. Three Chi-Square analyses were performed to determine the degree to which (a) the expected error categories matched the student responses, (b) the expected error categories matched the teacher perceptions, and (c) the student responses matched the teacher perceptions. The teacher ratings were then evaluated for their interrater agreement.

Finally, Chi-Square analyses were used to determine the association between student group membership (i.e., advanced 6^{th} grade versus general 7th grade mathematics and regular versus special education students) and the error categories. SPSS (IBM SPSS Version 22.0, 2013) was used to perform this analysis.

Validating a mathematics interim assessment with cognitively diagnostic error categories represents an important step in furthering the research for CDAs and interim assessments theoretically, empirically, and practically. First, the development of the error categories broadens the theoretical foundation of CDAs beyond skills, attributes, and levels of student understanding. Second, given that little empirical research has occurred with interim assessments, this study adds to the extant literature. Finally, this research provides a potentially meaningful method for improving student achievement in the present accountability culture.

Chapter II: Background

The goal of this study is to create and validate a mathematics interim assessment in which the test data informs instruction through a close connection between the test and student cognition. This chapter begins with a definition of "interim assessments" and then examines the purposes which drive interim assessment development. It subsequently reviews testing from two perspectives: interim assessments and their instructional use in schools versus cognitive diagnostic assessment models. Following is a brief look at cognition as it is related to mathematics. Finally, the chapter ends with a recommended test design that provides teachers with the necessary diagnostic data to better inform instruction.

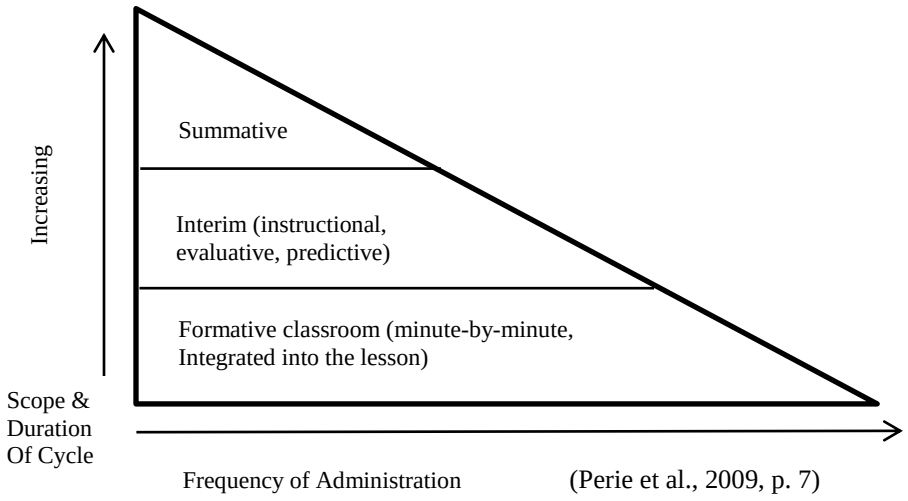
Interim Assessments Defined

Formative and summative assessments are well-known forms of assessments among educators, content specialists, school administrators, and test developers. Summative assessments are tests *of* learning in which the assessor is interested in a grade or score that indicates where the examinee is in their understanding of a set of concepts (Ainsworth & Viegut, 2006; Perie, Marion, & Gong, 2009). No remediation of skills is performed. Rather, after the score is recorded and explained, new learning occurs. Summative assessments have been administered for many years as classroom assessments and large-scale standardized assessments, as seen in the IOWA test of basic skills, Stanford 10, and state-level end-of-course (EOC) or end-of-grade (EOG) content tests like the Virginia SOL tests.

Formative assessments have a uniquely different goal from summative assessments. Formative assessments are "tests *for* learning" (Ainsworth & Viegut, 2006,

p. 23). With formative assessments, assessors seek to provide immediate feedback to the examinee and educator (Ainsworth & Viegut, 2006; Black & Wiliam, 1998ab; Wiliam, 2011). Formative assessments occur in the middle of the learning sequence, because the goal of formative assessments is to *inform* instruction. If the ensuing instruction is not modified to accommodate students' instructional needs, the assessment does not qualify as a formative assessment despite the intended purpose of the test (Black, Harrison, Lee, Marshall, & Wiliam, 2004; Black & Wiliam, 1998ab; Wiliam, 2011). Black and Wiliam (1998a) performed a meta-analysis of 250 formative assessment studies in which their research suggests that formative assessments are the most powerful type of assessment to improve teaching and student achievement.

Figure 1: Assessment tiers



(Perie et al., 2009, p. 7)

As seen in Figure 1, interim assessments lie characteristically *between* formative and summative assessments (Perie et al., 2009; Perie, Marion, Gong, & Wurtzel, 2007). Unlike formative assessments, interim assessments are not classroom assessments (Clune & White, 2008; Perie et al., 2009; Perie, Marion, Gong, & Wurtzel, 2007). They are

standardized assessments developed by testing companies, school district content specialist(s), or a combination of teachers and district level content specialist(s). Although teachers may participate in the development of interim assessments, they are typically not the sole authors. Whereas formative assessments are administered during or immediately following a specific curricular unit, interim assessments are only administered one to three times a year. As such, feedback from interim assessments is less immediate to instruction. Because interim assessments are administered less frequently, they can examine student retention of key content and student growth. Results from interim assessments are typically aggregated at the school or district level. School districts may have many reasons to implement interim assessments, the primary reason being to improve student achievement.

Purposes of Interim Assessments

Perie, Marion, and Gong (2009) identified three purposes of interim assessments: evaluative, predictive, and instructional. Interim assessments used for evaluative purposes are designed to compare teacher effectiveness, pedagogical techniques, or curriculum programs over time by aggregating student performance data (Perie et al., 2009). Evaluative interim assessment results may be aggregated over the course of one or more academic years. For instance, district level policymakers can use the data from evaluative interim assessments to inform the development of pacing guides based on the needs of large groups of students. Thus, evaluative interim assessments are focused on adapting curricular units, pacing guides, and pedagogical techniques for future students, not the students actually being assessed. An evaluative use of interim assessments requires standardization across schools. Should each school within a district develop and

administer different interim assessments, the assessments could not be aggregated, at which point these tests are no more helpful than a typical classroom assessment. Perie et al. (2009) recommend that evaluative interim assessments should (a) be aligned with the state's content standards, (b) contain items with a range of difficulty, and (c) comprise mixed item types. They believe these criteria would allow for a clearer understanding of a district's educational programs.

On the other hand, interim assessments used for predictive purposes are designed to determine a student's probability of achieving a criterion score on an EOC or EOG summative assessment (Perie et al., 2009). Given NCLB and many state's mandated requirements for student achievement, predictive interim assessments are an enticing goal for school districts to adopt. Should a school district elect to create predictive interim assessments, they could potentially target their remediation efforts to students who are predicted to fail the EOC or EOG summative assessment. To ensure the predictive capability of the interim assessments, Perie et al. (2009) recommend the interim assessment would need to be highly correlated with the criterion measure and test blueprint. In addition, both assessments should be composed of similar item types and difficulty levels. Although predicting a student's test score is valuable, most educators are likely interested in *how to improve* test scores of students who are predicted to fail EOC assessments (Perie et al., 2009). To do so requires an instructional use of interim assessments.

Interim assessments that are designed for instructional purposes are the most closely aligned with formative assessments because their focus is to modify instruction and improve student learning (Perie et al., 2009). Thus, instructional interim assessments

are aligned with state content standards allowing school districts to aggregate students' strengths and weaknesses according to those standards. Perie et al. (2009) contend teacher knowledge of students' strengths and weaknesses is not sufficient to modify instruction. Teachers need strategies for effectively altering their teaching to meet students' instructional and learning needs. Perie et al. (2009, p. 6) recommend that instructional interim assessments should fit "seamlessly with instruction" providing not only important data for educators but also opportunities for student learning. Finally, Perie et al. (2009) argue that instructional interim assessments should be developed so that educators have a clearer understanding of student cognition from test items' incorrect answers.

Even though each of these purposes is presented as distinctly separate, there is overlap. For example, if an interim assessment is used for instructional purposes, it *can* also be used for predictive purposes (Perie et al., 2009). Despite this apparent versatility in test design purposes, Perie et al. (2009) recommend that each interim assessment have a *primary* purpose ensuring that a test's goal is adequately and sufficiently fulfilled.

Henceforth, this research study will address interim assessments used for instructional purposes. For interim assessments to be truly instructional, the accompanying scores must yield diagnostic information which aids teachers in their understanding of student content strengths and weaknesses. Only then can teachers have the information with which they can remediate student misconceptions or extend student understandings.

Interim Assessments and Schools

Given the rigorous demands of NCLB and state content standards, such as the Virginia Standards of Learning (SOLs) and the Common Core Standards, teaching in public education is more and more concerned with accountability and student achievement (Little, 2012). Many policymakers once believed that EOC and EOG summative assessments would render diagnostically helpful information in improving student achievement (Perie et al., 2009), but this has proved false. These conditions have bread a market for interim assessments despite the paucity of research to support this venture (Goertz, Olah, & Riggan, 2009). In fact, within the testing world interim assessments are a relatively new phenomenon. The research literature describes interim assessments using many terms other than *interim assessments*, including benchmark assessments, periodic assessments, and formative assessments (Perie et al., 2009). Given the limited interim assessment research "it is unclear whether interim assessments would function as a system of classroom assessment capable of producing the major gains in student achievement attributed to formative classroom assessments" (Clune & White, 2008, p. 14).

The school reform, data-driven decision making (DDDM), and school improvement plan (SIP) research literature is replete with the *use* of interim assessments (see Carlson, Borman, & Robinson, 2011; Christman et al., 2009; Clune & White, 2008; Datnow, Park, & Wohlstetter, 2007; Goertz, Olah, & Riggan, 2009; Henderson, Petrosino, Guckenburg, & Hamilton, 2007; Marsh, Pane, & Hamilton, 2006; Olah, Lawrence, & Riggan, 2010), yet there is conflicting evidence that interim assessments are actually

effective in increasing student achievement (Christman et al., 2009; Goertz et al., 2009; Olah, Lawrence, & Riggan, 2010).

For instance, Henderson et al.'s (2007) quasi-experimental study used an interrupted time series to examine the impact of quarterly interim assessments on middle school mathematics achievement. They found no statistically significant or substantively important difference between the intervention (i.e., those schools implementing interim assessments) and control schools. The lack of a statistical difference could have occurred for several reasons: the study used school level data rather than student level data, data were lacking on what interim assessment practices existed in the control schools, and the researchers did not disaggregate the data by mathematics strand which might have made the results more sensitive to the interim assessments.

In the Clune and White (2008) qualitative study, the researchers investigated if the intended purposes of the interim assessments were achieved. These purposes included: greater alignment of the curriculum to state standards, additional practice for the state EOC or EOG test, and the generation of data to be used for instructional improvement. Although each of these purposes was achieved on some level, teachers complained most about the loss of instructional time. Furthermore, it was not clear how much instruction actually improved and if the improvements were sufficient to warrant the cost of the interim assessments. After three years the school district opted to cease using interim assessments.

In contrast, the Carlson et al. (2011) study is one of the first large-scale efforts to assess causal effects of a data-driven reform on achievement outcomes. The Carlson study used a multi-state district-level cluster randomized trial in over 500 schools within

59 school districts. The majority of the schools were low-performing, but diverse in their setting (e.g., urban vs. rural), ethnicity, and socio-economic status. The treatment group implemented interim assessments for the three years of the study, while the control group received a one year delay in the treatment. The interim assessments were created to mirror the state assessment blueprint and question types and to serve a predictive purpose. Correlation with the state test was between 0.80 and 0.85. During the first cohort year approximately 60-70% of the assessments were administered one to two times during the school year. Over the final two years, greater than 90% of the assessments were administered quarterly, which was in accordance with the test implementation design. With a mathematics effect size of approximately 0.20, the Carlson (2011) study is the best evidence that interim assessments can contribute substantively and significantly to improving mathematics achievement.

Despite the evidence from the Carlson study, it is just one study. There is no clear paradigm of data-driven decision making for school divisions to follow, especially with respect to interim assessments. With some studies, the intent is for the interim assessments to be used diagnostically, but some teachers end up using them predictively (Blanc et al., 2010; Christman et al., 2009). In other studies, the focus is to modify instruction based on the data (Christman et al., 2009; Clune & White, 2008; Goertz et al., 2009).

Instructional modification might refer to small-group instruction or peer teaching of "bubble students" who are just below the passing criterion score (Marsh et al., 2006). More often than not, teachers re-teach the procedural steps or go over the entire test demonstrating how to do each mathematics problem (Christman et al., 2009; Goertz et

al., 2009; Olah, Lawrence, & Riggan, 2010; Shepard, 2010). These studies provide little evidence that teachers remediate students' conceptual knowledge. In the Goertz et al. (2009) study only 10-15% of the test items allowed for conceptual inferences to be drawn, which gives credence to the procedural diagnoses. Additionally, only 2-3 items out of 20 contained common student errors as distractors. If multiple choice items are not designed to encompass students' common errors, procedural and conceptual diagnoses, then it is plausible that they will provide few opportunities to illuminate student understanding (Goertz et al., 2009; Goren, 2010).

The research literature provides numerous suggestions for interim assessment design and analysis so that student achievement might be improved. The most salient ones include (a) design multiple choice items to be diagnostic: contain students' common errors, procedural and conceptual knowledge distractors (Goertz et al., 2009; Goren, 2010; Mandinach & Honey, 2008; Shepard, 2010), (b) design the assessments to be cumulative so that student retention and progress can be tracked across the academic year (Marshall, 2008), (c) create a robust "feedback system" in which educators not only determine which students need remediation/intervention but they also reflect on how they need to change their instruction to promote learning (Blanc et al., 2010), (d) assess the effectiveness of the remediation interventions (Christman et al., 2009), and (e) provide more professional development to teachers in the analysis and interpretation of test data (Boudett, Murnane, & City, 2005; Mandinach & Honey, 2008).

Researchers are not the only ones making recommendations for changes to interim assessments and their analysis. Teachers are as well. Huff and Goodman (2007, p. 24) state that "educators are demanding … that they receive instructionally relevant

results from any assessments in which their students are required to participate and that these assessments be sufficiently aligned with classroom practice to be of maximum instructional value." But, what does such an interim assessment framework look like? For an interim assessment to be of "maximum instructional value" it should be diagnostic so that teachers have sufficient feedback to determine not only student content strengths and weaknesses, but also indicates the processes students use in their problem solving (Boudett, City, & Murnane, 2010). However, caution should be exercised so that the amount of this data is not overwhelming to teachers. This raises another concern, data analysis.

Several researchers have noted that teachers as a group do not possess a high degree of assessment literacy or lack a background in statistics (Mandinach & Honey, 2008; Marsh et al., 2006; Young, 2006). Many of these same researchers have recommended that teachers have professional development in how to use assessment protocols to frame their data analysis (Bambrick-Santoyo, 2010; Boudett et al., 2005; Christman et al., 2009; Datnow et al., 2007).

A further concern is the type of items that comprise the interim assessment. Because of scoring time constraints most of the interim assessments in the aforementioned studies are multiple choice tests. Thus, for this study an interim assessment framework needs to be found that (a) is cognitively diagnostic, (b) is multiple choice, (c) has data that is not too small or large grain, and (d) illuminates the processes students use in their mathematics problem solving. One assessment framework that can potentially satisfy each of these conditions is cognitive diagnostic assessment models.

Cognitive Diagnostic Assessment Models

A cognitive diagnostic assessment (CDA) model is a model of student learning which diagnoses student "knowledge structures and cognitive processing skills" so that remediation is informed (Leighton & Gierl, 2007b, p. 3; Nichols, 1994). Leighton and Gierl (2007a) suggest that if a cognitive model is not empirically derived, then it cannot support diagnostic inferences. Nichols (1994) submits that the underlying cognitive theory of CDA is used to *generate* assessments and predict results. Assuming that a CDA meets both conditions set forth by Leighton, Gierl, and Nichols, the resulting data should allow teachers to "alter student misconceptions and faulty strategies" (Leighton & Gierl, 2007b, p. 6). However, despite these theoretical definitions, no clear testing framework for CDA achievement tests has been established.

I will review several CDA models to discern how suitable they are for a mathematics interim assessment. My goal is to find a CDA that is not too large grain that the test score data is not diagnostically helpful or too small grain that the amount of data is overwhelming to teachers. Therefore, the CDA model needs to have data which is (a) cognitively diagnostic, (b) "medium grain," and (c) practically useful to teachers in their instructional planning and remediation.

Gorin (2007) identified two types of CDAs: statistical models and cognitive models. Gorin defined statistical models as those which focus primarily on post hoc tests and item analysis, specifically the rule space model and the attribute hierarchy model. Gorin contends that cognitive models implement cognitive theory throughout the item and test development process. She cited Mislevy's Evidence-Centered design and Embretson's Cognitive Design system as examples of cognitive models. Lastly, Gorin

discussed ordered multiple choice questions as a potential CDA model given two significant changes to the typical multiple choice paradigm: the tests are not binarily scored and the distractors are written using student misconceptions.

Given Gorin's analysis of CDA assessments, I will begin with an examination of Tatsuoka's Rule-Space model, since Tatsuoka was "one of the first psychometricians to embrace the union" of cognitive psychology and psychometric theory (Gierl, Leighton, & Hunka, 2000, p. 41). Then, I will proceed with a discussion of Leighton and Gierl's Attribute Hierarchy Method, Mislevy's Evidence-Centered Design, Embretson's Cognitive Design System, and Ordered Multiple Choice categories. One other CDA model I will examine is the Diagnostic Classification Model which has received significant attention in the literature. Although other CDA models exist, I am focusing on those which have contributed substantially to the cognitive diagnostic assessment literature.

Rule Space model. During the 1980s binary scoring was the most common method of scoring achievement tests, but Tatsuoka (1983, 1986, 2009) believed that total scores lack valuable information about student's errors and misconceptions. For instance, two students with the same score of 50% may have missed different problems altogether or they may have missed the same problems for different reasons (Tatsuoka, 2009). Tatsuoka (1983) also argued that some students may select the correct answer using inaccurate rules or procedures. Thus, Tatsuoka (2009) reasoned total scores are not cognitively diagnostic and they do not provide sufficient information for remediation. To ameliorate the inherent problems with binary scoring, Tatsuoka (1983) developed the Rule-Space model (RSM). Her goal was to provide a testing framework which (a)

classified and diagnosed student *error* patterns and (b) provided a mechanism for instructional evaluation and student remediation (Tatsuoka, 1983, 1986, 2009).

The RSM is a deterministic and probabilistic model that combines error theory and item response theory (IRT) to classify and diagnose student's cognitive errors (Tatsuoka, 1983, 1986, 2009). The RSM is comprised of two stages, the selection of feature variables and statistical pattern classification.

The selection of feature variables stage encompasses the development of the Q matrix theory and the subsequent formation of several matrices: the adjacency matrix, the reachability matrix, the Q matrix, the reduced Q matrix, and the ideal attribute matrix (Gierl et al., 2000; Tatsuoka, 2009). Originally, these matrices were characterized by rules or procedures that examinees must perform to correctly answer a test item (Tatsuoka, 1983, 1986). For example, Tatsuoka (1983) described Rule 2 in one study as,

> the student uses a wrong rule for addition. He or she subtracts the smaller absolute value from the larger absolute value and takes the sign of the first number in the answer. The student converts subtraction to addition problems correctly, [and] then consistently applies the same erroneous rule to the new addition problem. (p.346)

With additional research, Tatsuoka (1995) broadened the matrices to comprise items and more general *attributes*, rather than only items and procedural rules. Tatsuoka defined attributes as unobservable knowledge, skills, or procedures necessary to correctly answer a test item (Birenbaum et al., 1993; Tatsuoka, 1995, 2009). The construction of attributes as models of item performance is hypothesized by content experts, such as cognitive researchers and educators (Birenbaum et al., 1993; Tatsuoka, 2009). Once identified, the attributes are ordered hierarchically, although this is not a requirement with the RSM (Gierl, 2007; Tatsuoka, 1990). In short, the matrices are created so that (a)

items can be matched with attributes at a fine grain and (b) cognitive diagnosis of student error patterns can be achieved (Birenbaum et al., 1993; Gierl, Alves, & Majeau, 2010; Gierl et al., 2000; Tatsuoka, 1983, 1986, 2009).

Following matrix development is the construction of the ideal item response vector, or ideal knowledge states. The ideal item response vector contains the ideal attributes paired with a total score. If many examinee score patterns occur that do not fit the ideal item response vector, then one or more of the following is true, (a) the attributes were not correctly identified, (b) the attribute hierarchy is not true, (c) the items did not coincide with the cognitive model, (d) the test was incongruous with the examinees sampled, or (e) several slips occurred (Gierl, et al., 2000). Once the ideal item response vector is complete, item generation begins.

The next stage, statistical pattern classification occurs within the rule space where examinee's item response patterns are plotted and compared to ideal item response patterns (Birenbaum et al., 1993). The rule space is a two dimensional Cartesian coordinate system where the x-axis is defined as theta, θ or ability and the y-axis is defined as zeta, ζ , the person-fit statistic or the unusualness of an item response (Birenbaum, et al., 1993; Gierl, 2007; Tatsuoka, 1983, 1986, 2009). The rule space is a statistical tool used for the classification of item response patterns by way of the examinee's mastery (and non-mastery) of attributes (Birenbaum et al., 1993). As such, an examinee's item response pattern forms a picture of their *likely* knowledge state. The ideal knowledge state or correct response is found at the point (1, 1). The ordered pair, (θ, ζ), indicates the distance from each of the ideal knowledge states. Points with higher θ values indicate more ability while lower θ values indicate less ability. In turn, points

further from the θ axis indicate students with more unusual responses while points closer to the x-axis represent students with more common item responses. In order to classify a student's knowledge state, a probability ellipse is drawn around each student's item responses. The shortest Mahalanobis distance between a student's item response pattern and the ideal item response pattern indicates the student's knowledge state and the most conservative attribute mastery pattern (Birenbaum et al., 1993). In practice, the shortest Mahalanobis distance means teachers are likely to capture all concepts and skills students have not mastered. Thus, the attribute mastery pattern provides a mechanism for instructional planning and remediation. However, because this method of identifying knowledge states and the attribute mastery patterns is based on probability, teachers may reteach some concepts students already know.

With many applications of the RSM, the attribute hierarchy cannot be fully developed because the RSM is *retrofitted* to an extant test (Gierl, 2007), although this is not always the case (Birenbaum et al., 1993). If retrofitting does occur, the test may not have the necessary item types to generate a Q matrix (Gierl, 2007). Without a Q matrix, the cognitive model is incomplete because the hierarchy is relatively simple and rule-based or there are "few hierarchical relationships explicitly represented between attributes" (Gierl, 2007, p. 334).

In response to the RSM, many other CDA models have been developed; one of these is the Attribute Hierarchy method (AHM).

Attribute Hierarchy method. Because the AHM is an extension of the RSM, many features of these models are the same. Thus, my discussion will focus on the key similarities and differences between both models.

Like the RSM, the AHM employs Q matrix theory and attributes which are developed by content experts (Gierl, 2007; Gierl, Leighton, & Hunka, 2007; Leighton, Gierl, & Hunka, 2004). AHM attributes are represented in the same type of matrices seen in Tatsuoka's RSM: the adjacency matrix, the reachability matrix, the Q matrix, the reduced Q matrix, and the ideal attribute matrix. The fundamental difference between the RSM and the AHM is in their assumptions about *how* cognitive attributes are modeled (Leighton, Gierl, & Hunka, 2004). The RSM does not require that attributes are hierarchically ordered (Tatsuoka, 1990), although the AHM does (Leighton et al., 2004). Because an attribute hierarchy presupposes dependence among the cognitive attributes, the attributes with the AHM are dependent. Since the RSM lacks a *required* attribute hierarchy, the attributes may be independent or dependent. Furthermore, with the AHM four hierarchical attribute structures are employed to demonstrate the direct and indirect relations which can exist between attributes (Gierl et al., 2007; Leighton et al., 2004). These hierarchical attribute structures can be used alone or combined into more complex structures. No such attribute structure exists with the RSM.

As previously mentioned the RSM is often *retrofitted* to an existing test which forces a Q matrix upon test items not designed for cognitive diagnosis (Gierl, 2007). The resulting cognitive model is developed post hoc, which compromises the cognitive skill capability of the assessment. On the other hand, the AHM depends on an *a priori* development of the attribute hierarchy (Leighton et al., 2004). This process supports (a) an attribute hierarchy that guides item development, (b) a more fully developed Q matrix, and (c) a greater capability for cognitive skill diagnosis.

Finally, in both models statistical pattern classification occurs within the rule space by comparing observed examinee response patterns to ideal response patterns. With the RSM, the purpose of statistical pattern classification is to determine examinees' probable cognitive *errors* and *misconceptions* (Gierl, 2007; Tatsuoka, 1990). However, the goal in the AHM is to discern which attributes examinees are likely to have *mastered* (Gierl, 2007).

I propose that if a test design focuses on cognitive strengths/content mastery, then the teacher lacks sufficient information in how to remediate student weaknesses. When students answer an item incorrectly, they are all not likely to have arrived at their wrong response in the same way, or for the same reasons. Error analysis information has the potential to better inform instruction if it points to ways in which examinees responded incorrectly.

Diagnostic Classification Model. The DCM is a statistical model that *predicts* student performance according to a set of mastered attributes (Rupp & Templin, 2008; Rupp et al., 2010). Like AHM, the DCM employs a Q matrix to specify the item attribute relationship where attributes are assigned a priori. However, because the DCM allows any given item to have more than one latent skill load on it, the items are multidimensional. For instance, if one attribute is addition and another is subtraction, one item might assess both the addition and subtraction attributes.

The optimal test purpose for DCMs is as a classification measure, such as with EOG/EOC tests. In this type of assessment students are classified as pass or fail. More often they are divided into achievement levels, such as pass advanced, pass proficient, basic, and below basic.

Unlike most assessment models, the DCM does not assign a scaled score, but rather a profile of mastered skills or attributes. This profile of mastered skills is obtained by comparing students' observed behaviors to two models, a measurement model and a structural model. Measurement model item parameters indicate how students with different diagnostic profiles respond to a group of test items. These results reveal which items are better at distinguishing between students with dissimilar diagnostic profiles. The structural model compares the frequency of the student diagnostic profiles in the population. These results help validate the credibility of the observed diagnostic profiles. The resulting student mastery profiles are presented as *probabilities* of mastery. Despite the attribute mastery profiles, DCMs focus on why a student is not performing well. Because this model is centered on skills acquired rather than on a cognitive processing diagnosis, it is not an appropriate model for this study.

Evidence-Centered Design. Mislevy's Evidence-Centered Design (ECD) is a detailed, comprehensive assessment design framework involving a series of evidentiary arguments (Behrens, Mislevy, Bauer, Williamson, & Levy, 2004; Gorin, 2007; Mislevy, Almond, & Lukas, 2003; Mislevy & Haertel, 2006; Williamson, Bauer, Steinberg, Mislevy, Behrens, & DeMark, 2004). Mislevy begins by breaking down the test development process into five layers (Mislevy & Haertel, 2006) and four models (Gorin, 2007; Mislevy et al., 2003; Williamson et al., 2004). Each layer and model is labeled using a unique terminology not seen with other test design frameworks. Together these layers and models address the design, implementation, and delivery of an educational assessment. Mislevy argues that ECD's structural design creates a common language among professionals engaged at varying facets of the test development process (Behrens

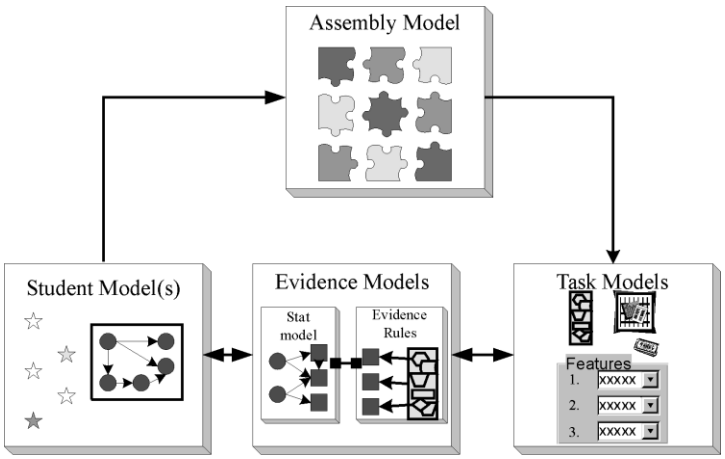
et al., 2004; Mislevy et al., 2003). He further adds that the "consistent use of the terminology of ECD [gives] us a language to conceptualize and articulate possibilities that could not have been brokered with older language" (Behrens et al., 2004, p. 299).

First, I will briefly discuss ECD's five layers, which encompass domain analysis, domain modeling, conceptual assessment framework, assessment implementation, and assessment delivery (Mislevy & Haertel, 2006). Because the four model structure is subsumed within ECD's conceptual assessment framework (CAF) layer, I will briefly examine its structure during the CAF discussion.

The first layer of ECD test development is domain analysis which is the collection of content area information to be assessed. This content information may consist of state standards, concept maps, domain specific terminology, knowledge representations, tools, and domain specific notation (Mislevy & Haertel, 2006). Subsequently, in domain modeling content experts focus on the "*big ideas of a given domain*" (Mislevy & Haertel, 2006, p. 8). Here content experts decide what the test will measure and how it will be executed. Examples of domain modeling include assessment argument diagrams, potential observations or rubrics, potential work products, and primary knowledge and skills.

The third layer, the conceptual assessment framework (CAF) is the backbone or *blueprint* for the assessment (Mislevy et al., 2003). The CAF encompasses the technical specifications, evaluation procedures, and measurement models of the test. The CAF is organized around four models: the student model, evidence model, task model, and the assembly model (Gorin, 2007; Mislevy et al., 2003; Mislevy & Haertel, 2006; Williamson et al., 2004) (See Figure 2).

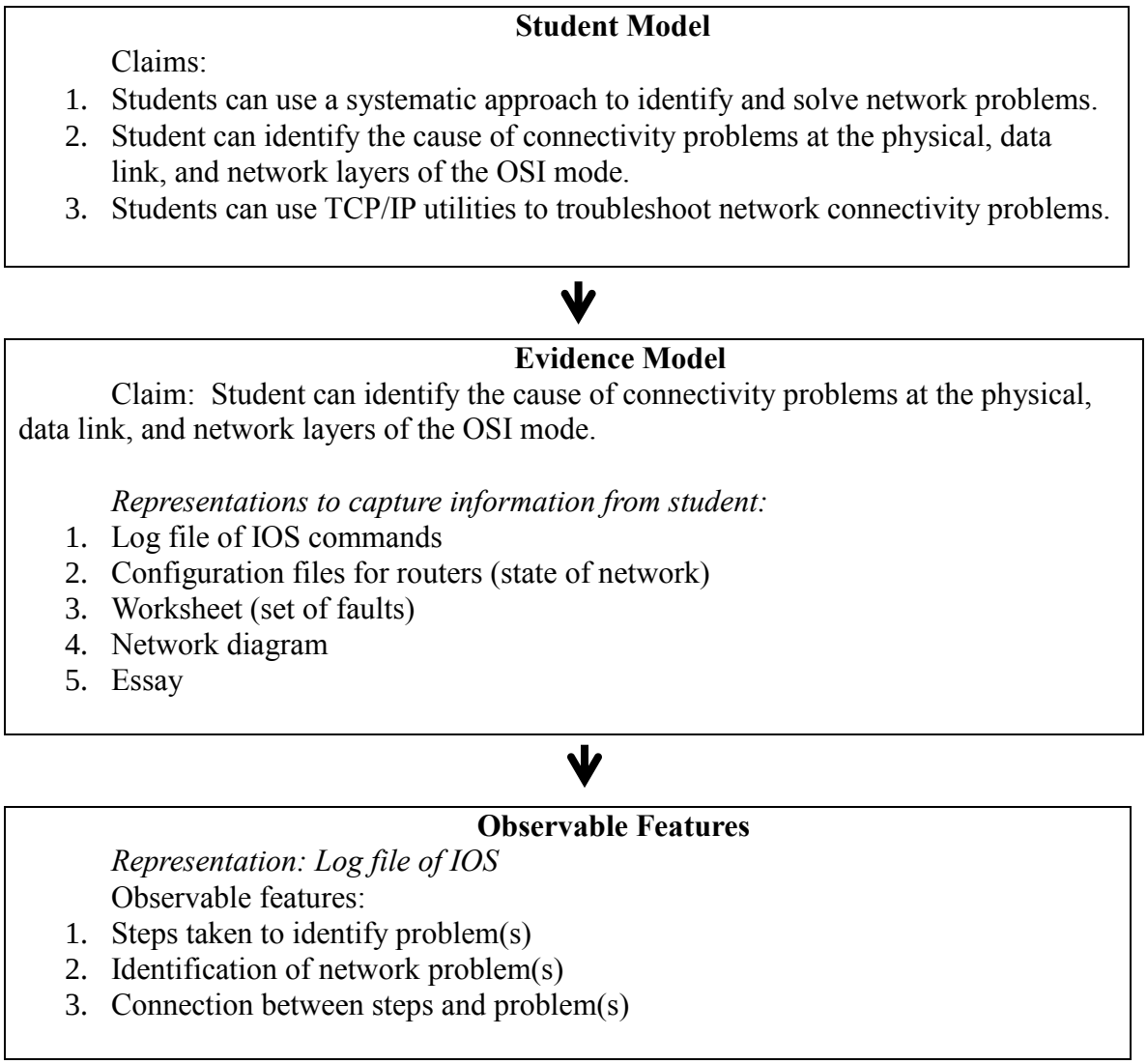
Figure 2: The four central models of an ECD framework



Williamson et al., 2004, p. 306

The student model is represented by a set of claims pertaining to the knowledge, skills, or abilities the test seeks to measure (Gorin, 2007; Mislevy et al., 2003; Mislevy & Haertel, 2006; Williamson et al., 2004). For instance, in the NetPASS computer networking assessment, one of the student model claims was "students can identify the cause of connectivity problems at the physical, data link, and network layers of the OSI mode" (Gorin, 2007, p. 180; Williamson et al., 2004, p. 311) (See Figure 3). Because the NetPASS claim uses the language, "*students can identify…,"* the claim is measurable. State content standards in many states are written in similar *measurable* ways as seen in the NetPASS claim example. For instance, one of the 7th grade mathematics Virginia SOL standards, 7.15b is the student will graph solutions to inequalities on a number line (Virginia Department of Education [VDOE], 2009, Curriculum Framework: 7th grade, p. 26). Virginia SOL standard 7.15b is measurable because the assessor can gather evidence to demonstrate mastery of this skill. Perhaps, then, an abbreviated definition of a student model is a set of *measurable* content standards or claims.

Figure 3: Claim and Evidence Chain from the NetPASS assessment



(Gorin, 2007, p. 180)

The evidence model provides evidentiary representations in support of the student model claims (Gorin, 2007; Mislevy et al., 2003; Mislevy & Haertel, 2006; Williamson et al., 2004). The evidence model gathers this data through two components: evidence rules and a measurement model. Evidence rules guide the item response scoring process because they govern the identification and summary of evidence *within* tasks. Continuing with the NetPASS example, there are several proposed forms of evidence that could demonstrate mastery of the aforementioned claim. Evidence could manifest as a

log file of IOS commands, configuration files for routers, a worksheet, a network diagram, or an essay (Gorin, 2007, p. 180; Williamson et al., 2004, p. 311). In contrast, the measurement model guides the summary scoring process because it is concerned with the accumulation and synthesis of evidence *across* tasks (Mislevy et al., 2003; Williamson et al., 2004). Measurement models consist primarily of psychometric models such as classical test theory, item response theory, and cognitive diagnostic models.

The task model describes the observable features of each evidence model representation (Gorin, 2007; Mislevy et al., 2003; Mislevy & Haertel, 2006; Williamson et al., 2004). Again using the NetPASS example, the observable features represented in the task model include (a) steps taken to identify problems, (b) identification of network problems, and (c) connection between step and problems (Gorin, 2007, p. 180; Williamson et al., 2004, p. 311). A typical assessment has more than one task model. In fact, a task model often comprises a "family of potential tasks" with different item types requiring different task models (Mislevy et al., 2003, p. 11).

Finally, the assembly model expresses how the student, evidence, and task models work together to present the final assessment (Mislevy et al., 2003). Assembly rules dictate a host of decisions such as the content of reading passages, the number of items that use mathematical figures, sentence complexity, and the mix of item types. Overall, the assembly model explains how the tasks are organized and presented to best represent the domain being assessed. Taken together the student, evidence, task, and assembly models provide an evidentiary chain that supports inferences about students' mastery of specific skills. This organizational model structure confirms why the CAF is seen as the blueprint of the ECD framework.

The last two layers, assessment implementation and assessment delivery pertain to the implementation and scoring of the tests, respectively (Mislevy & Haertel, 2006). Assessment implementation includes such activities as fitting the measurement models to the test scores, using piloted test data to refine evaluation procedures, and the development of task materials and tools. In turn, assessment delivery entails examinee's interaction with tasks, task and test-level scoring, and score reporting. Examples include numerical and graphical summaries of test scores by examinee and by group.

ECD is a construct-centered approach that has been criticized by some researchers as merely a renaming of test development procedures and standards that have been implemented for many years (Mislevy & Haertel, 2006). Mislevy recognizes that "all of the innovations… [such as] in cognitive psychology, …measurement models, task design, scoring methods…have been developed by thousands of researchers across many fields of study, without particular regard for ECD" (p. 17). However, Mislevy contends that ECD is

> a framework that does indeed provide new words for things we are already doing. But, it helps us understand what we are doing at a more fundamental level. And it sets the stage for doing what we do now more efficiently, and learning more quickly how to assess in ways that we do not do now (p. 18).

Despite Mislevy's contentions, the ECD does not appear to offer significant differences from more traditional models in informing teachers about students' cognitive processes relative to their mathematics problem solving. Susan Embretson's Cognitive Design System offers a different perspective on how to merge cognitive theory with test design and item development. A brief description of her model follows.

Cognitive Design System. Embretson's Cognitive Design System (CDS) is an assessment model in which cognitive theory occupies a central role in test design and item development (Embretson, 1998, 1999, 2010; Embretson & Gorin, 2001; Gorin, 2007). Unlike more traditional cognitive models, with the CDS cognitive theory *precedes* test design and item development. This cognitive theory focus is managed through two frameworks, a conceptual framework and a procedural framework.

The conceptual framework addresses the construct validation of a measure through two components: construct representation and nomothetic span (Embretson, 1983, 1998; Embretson & Gorin, 2001). Construct representation involves the "meaning of test scores" (Embretson, 1998, p. 382). This *meaning* is established by understanding the processes, skills, and knowledge examinees use in their problem solving and then subsequently linking item features to cognitive theory. As a result "test items can be designed to reflect *specified* cognitive constructs" (Embretson, 1998, p. 382). While construct representation is concerned with the meaning of test scores, nomothetic span is concerned with the usefulness of test scores for measuring individual differences. Nomothetic span is assessed by examining the correlations of test scores with the "strength, frequency and pattern of significant relations with other measures" (Embretson, 1983, p. 180). In short, construct representation is about "identifying the theoretical mechanisms that underlie task performance" while nomothetic span is about the relationship of a test to other external measures (Embretson, 1983, p. 180).

The procedural framework is comprised of a series of *iterative* stages where cognitive theory drives test design, item development, and the estimation of item parameters (Embretson, 1998, 2001). Through these stages, the procedural framework

provides a process model of item performance and it relates item processes to test validity. The procedural framework includes seven stages: (a) specify the goals of measurement, (b) identify design features in the task domain, (c) develop the cognitive model, (d) generate items, (e) evaluate models for generated tests, (f) bank items by cognitive complexity, and (g) validation: nomothetic span. A brief description of each stage follows.

In the first stage, *specify the goals of measurement*, the goals for construct representation and nomothetic span are articulated. With the second stage, *identify design features in the task domain*, potential item types are reviewed in the cognitive psychology literature. In addition, differing task features are considered with respect to process, skill, and knowledge structures and a test's measurement goals. Although these first two stages are typically seen in traditional test development, with CDS cognitive theory plays a significant role.

The next two stages are essential to the CDS because this is where the construct representation of the test is explicitly defined and construct validity is elaborated. The third stage is the *development of the cognitive model.* Here "relevant cognitive processes, strategies, and knowledge structures must be identified and organized into a unified model" (Embretson & Gorin, 2001, p. 351). Then, this model must be integrated with the cognitive psychology research and the chosen item type(s). For each item type, item features must be explicitly manipulable, defined, and scoreable so that the differing cognitive processes can be represented in test items. Next, the cognitive item features must be studied empirically to determine the impact on the psychometric properties of the test. Once the cognitive model is firmly established, then in the fourth stage the cognitive

model is operationalized to generate test items. Ideally, the differences in test item structures should represent differences in cognitive processes.

The last stages elucidate the psychometric properties of the assessment. In the fifth stage the cognitive and psychometric models must be evaluated to determine the fit of the cognitive theory to the item response data. Cognitive model fit is established by predicting item performance. The independent variables encompass item structures and item stimulus features, while the dependent variables often include response time and item difficulty. The impact of the item structures and stimulus features indicates the relative cognitive processes these items represent. In contrast, the psychometric model fit is an evaluation of the fit of an IRT model to the item response data. Should the IRT model not sufficiently fit the item response data, then the cognitive model needs to be revised. Poor psychometric model fit could occur for two reasons: convergent or divergent data. For example, if manipulations of the item variables create no change, or an unpredictable change in the item parameters, then the model does not sufficiently account for the relationship between the item features and the cognitive model. Or, if variables not included in the model affect the item parameters, then the model is lacking important variables which account for significant sources of variance. By testing and manipulating item features, stronger conclusions can be made about the effects of the cognitive model on test scores. Subsequently, in the sixth stage, *bank items by cognitive complexity,* test items can be stored in a test bank according to their cognitive complexity. This process assumes, of course, that the cognitive and psychometric models fit the test data well. If the psychometric model sufficiently predicts item performance, then items can be apportioned according to features that contribute toward item difficulty. Items can

then be categorized according to their difficulty and cognitive complexity. Finally, the last stage is *validation or nomothetic span*. Here test items are evaluated as to how well they correlate with other external measures of the construct(s).

Embretson and Gorin (2001) contend that the CDS approach renders many advantages over other CDA models:

> (1) Item parameters may be predicted for newly developed items.
>
> (2) Construct validity is more completely understood.
>
> (3) Construct validity may be understood at the item level.
>
> (4) Enhanced score interpretations are feasible if IRT scaling is used.
>
> (5) Items may be developed for specified sources of cognitive complexity.
>
> (6) Computer generation of items with specified sources and levels of item difficulty may be feasible (p. 352).

Despite the union of cognitive theory and testing, Embretson believes that the "most important potential for cognitive theory is test design" (Embretson & Gorin, 2001, p. 365). Embretson contends that traditional construct validity paradigms relegate cognitive theory as a post-hoc interpretation after a test design is complete. Embretson insists that cognitive theory can have a profound impact on construct representation if it is used prior to test design and development. For instance, Embretson has used CDS for ability and achievement tests, although most often with ability measures. In an object assembly test, Embretson used a four distractor multiple choice model to define her item structure. Her most prominent finding from the object assembly cognitive model was that the decision process was impacted by the nature of the distractors (Embretson & Gorin, 2001, p. 360). Thus far, no other CDA has described using a multiple-choice item

structure, yet this item structure is immensely popular with achievement tests. Additionally, none of the other CDA models have discussed the significance that multiple choice distractors can play in the decision process of mathematical problem solving. The next model will not only explore a multiple-choice assessment model, but it will also examine cognitive diagnosis from the perspective of the item distractors.

Ordered Multiple-Choice tests. Traditional multiple-choice (TMC) tests are often viewed as assessing a low level of content and cognitive demand, especially definitions and the recall of facts (Hamilton, Nussbaum, & Snow, 1997). Scoring is dichotomous with responses scored as correct or incorrect. The resulting scores describe which item content students answered correctly, but it reveals nothing about students' incorrect responses. As such, dichotomous scoring loses valuable data about student misconceptions making instructional modification difficult (Black & Wiliam, 1998b) and fueling criticisms for why multiple-choice assessments cannot be used for cognitive diagnosis (Hermann-Abell & DeBoer, 2011).

In response to TMC assessments, Briggs et al. (2006) have devised ordered multiple-choice (OMC) assessments in which item distractors are written using common student misconceptions and linked to students' developmental levels (Haladyna & Rodriguez, 2013). OMC assessments blend the objectivity of TMC assessments with the richness of constructed response assessments (Briggs & Alonzo, 2009).

In OMC assessments response options are ordered along a progress variable (Wilson, 2008). Wilson (2008) defines a progress variable as a continuum of content understanding divided into successive levels of development. The highest level exemplifies complete understanding of a concept while the lowest level represents the

least understanding. Progress variables are derived iteratively from the research literature and a blend of professional opinion about what demonstrates higher and lower levels of understanding. But, as Alonzo and Steedle (2009) point out there may not be enough research to depict each level of the progress variable, or enough research to determine the relationship between levels. Therefore, progress variables are a "hypothesis about student thinking, rather than a description" (Alonzo & Steedle, 2009, p. 393). Once the progress variables are developed they are submitted to a series of validation procedures (e.g., cognitive think-alouds and clinical interviews).

Progress variables are often depicted in a construct map (Wilson, 2005). Misconceptions represented at one level of a construct map are resolved at the next level (Wilson, 2008). Table 1 depicts a construct map for the properties of light. Each level represents the knowledge that a student possesses and possible misconceptions or common errors. Misconceptions at each level should not be thought of merely as errors but as developmental stages (Briggs, Alonzo, Schwab, & Wilson, 2006). Score level 4 indicates that the student completely understand the properties of light while score level 1 demonstrates the least understanding. A construct map can be especially helpful to teachers in their instructional modification and differentiation. Some construct maps depict common error(s) more explicitly than seen in Table 1 (Briggs et al., 2006, p. 42).

Table 1
A Construct Map of the Properties of Light

Score Level	Description
4	Student conceives of light as a distinct entity in space. Understands the relationship between a light's source, its motion and path, the objects it encounters along the way and the effect it produces.
3	Student conceives of light as a distinct entity in space, traveling in a straight line. Lacks an understanding of how light interacts with objects.
2	Student understands limited cause and effect relationships between a light's source (bulb), state (brightness) and the effect it produces (patch of light).
1	Student identifies light solely with respect to its source or its effect. Light is not understood apart from its effects. Student defines light in relation to dark.

Briggs et al., 2006, p. 38

Some of the research literature uses the term *learning progression* instead of progress variables, especially the science education literature (Alonzo & Steedle, 2009). In the mathematics education literature, learning progressions are called *learning trajectories* (Daro, Mosher, & Corcoran, 2011). But, the term learning progression/learning trajectory does not have the same meaning to all researchers or content areas. For example, some researchers use the term learning progressions/trajectories to represent broader learning over several years, such as when students learn multiplicative thinking or rational number reasoning. Daro et al. (2011) state this trajectory occurs from kindergarten through 8th grade. In this paper learning progressions/trajectories refer to student thinking over a *curricular unit*. Thus, in this context the terms progress variables and learning progressions/trajectories are interchangeable.

Psychometric modeling of OMC assessments is often achieved using item response theory (IRT), especially the Rasch model (Hermann-Abell & DeBoer, 2011), Partial Credit Model or the Ordered Partition Model (Briggs et al., 2006; Wilson, 2008). However, Briggs and Alonzo (2009, p. 8) find several potential problems with construct maps and IRT models. They argue that in OMC construct maps (a) the location of category thresholds is inconsistent, (b) cutpoints cannot be used to classify students into specific categories because student ability and item category thresholds are estimated with error. These standard errors of measurement are largest at the highest and lowest ends of the logit continuum, (c) a continuous interpretation of achievement is suggested yet student achievement levels are ordinal, and (d) multiple response options for some items are placed at the same level. In fact, Briggs and Alonzo posit that no model in the Rasch family of IRT models can solve these problems. Their solution is the Attribute Hierarchy Method (AHM) which they contend pushes the test developer to explicitly define not only the attributes but also the movement from one skill level to the next. However, the AHM is not without its problems. First, the AHM is primarily used for very fine-grained diagnoses, which does not fit with OMC assessments. Since most interim assessments are multiple-choice tests with three to four distractors and consist of approximately 20-30 items, the average multiple-choice test cannot handle a fine-grain diagnosis and still cover the necessary curricula. Second, Briggs and Alonzo (2009) assert that if a progress variable is more qualitative in its orientation then it is less conducive to the AHM. Lastly, the AHM has no model fit indices, yet IRT models do. Given the aforementioned conditions, there appears to be little benefit to using the AHM for the psychometric modeling of OMC assessments.

According to Alonzo and Steedle (2009), scoring of OMC assessments reveals that expert students read an item and then they categorize the problem according to rules and principles. In contrast, novice students pay more attention to surface features of the problems. This practice illuminates how novice students might display misconceptions in their item responses. Alonzo and Steedle (2009) also discovered that students' vocabulary deepens in concert with their conceptual understanding. Both of these distinctions between expert and novice students provide guidance for writing OMC items.

Finally, OMC assessments offer two advantages as a framework for interim assessments: they maintain the reliability advantages of TMC scoring even though OMC assessments have polytomous items (Briggs et al., 2006) and they offer teachers rich diagnostic information without being overwhelming. Thus far, OMC assessments have become increasingly popular in science education (Alonzo & Steedle, 2009; Briggs & Alonzo, 2009; Hermann-Abell & DeBoer, 2011; Wilson, 2008) but, no comparable OMC assessments have been devised in mathematics.

A Recommended Test Design

Several of the CDA models claim that they are employing an "information processing approach" in their assessment design or cognitive skills diagnosis (Gierl et al., 2007). Information processing theories usually contain components such as declarative knowledge, procedural knowledge, attention, processing speed, and working memory (Baddeley, 2007; Bjorklund, 2005; Dehn, 2008; Matlin, 2002). Broadly speaking, information processing theories deal with input (e.g., thinking processes) and output (e.g., skills) components. The aforementioned CDA models do not explicitly reference input

components instead they focus exclusively on students' skills, which are output components.

Although understanding students' skill levels is important, teachers in this current accountability climate are demanding "they receive instructionally relevant results from any [required] assessments… and that these assessments be sufficiently aligned with classroom practice to be of maximum instructional value" (Huff & Goodman, 2007, p. 24). Most teachers believe the diagnostic information currently available in most large-scale score reports is not detailed enough (Huff & Goodman, 2007, p. 44). Moreover, over 80% of the teachers surveyed by Huff and Goodman felt it was important to have suggested instructional strategies to accompany student diagnostic data. Despite teacher demands most interim assessments are multiple-choice tests where item responses are scored dichotomously. The resulting scores describe students' correct responses, but reveal nothing about students' incorrect responses. As such, these assessments lose valuable data about student misconceptions making instructional modification difficult (Black & Wiliam, 1998a) and fueling criticisms for why they cannot be used for cognitive diagnosis (Hermann-Abell & DeBoer, 2011). Furthermore, some CDA models are fine-grain tests that potentially inundate teachers with too much data. A balance needs to be found so that teachers have meaningful data that is not overwhelming. If interim assessments could point specifically to where student misconceptions lie and where deficits are in student's cognitive processes, interim assessments could foreseeably render "maximum instructional value." My goal is to develop an interim assessment that resolves each of these challenges.

Inspiration for how to develop an interim assessment to meet each of these challenges came from the cognitive and interim assessment literature (Baddeley, 2007; Bjorklund, 2005; Dehn, 2008; Feifer & De Fina, 2005; Goertz et al., 2009; Matlin, 2002; Mazzocco & Devlin, 2008). For instance, Goertz et al. (2009) investigated how teachers used multiple-choice interim assessments to modify instruction. These teachers developed four error categories to explain student performance: a *procedural-conceptual continuum, conceptual understanding, other cognitive weaknesses* which included test anxiety, difficulty maintaining attention, and weak reading ability, and *contextual diagnoses* which were outside the realm of school influence.

To capture the spirit of the aforementioned categories and provide a mechanism for teachers to more easily differentiate and remediate instruction, I redefined the error categories as *procedural knowledge, conceptual* or *declarative knowledge,* and *attention.* As such, these error categories fit with an information processing approach to cognition (Baddeley, 2007; Bjorklund, 2005; Dehn, 2008; Feifer & De Fina, 2005; Matlin, 2002) as well as the five interwoven strands of mathematical proficiency espoused by the National Research Council (NRC, 2001). The NRC's strands of mathematical proficiency encompass conceptual understanding, procedural fluency, strategic competence, adaptive reasoning, and productive disposition. The succeeding paragraphs will briefly examine cognition and these new error categories (attention, procedural knowledge, and conceptual knowledge) from the perspective of mathematical problem solving.

Mathematics cognition and problem solving. Matlin (2002, p. 364) contends that "attention is a necessary *initial* [emphasis added] component of understanding a

problem." Based on Matlin's statement, attention is required *before* problem solving can begin. But, attention is not a solitary construct.

Attention is an important feature of Baddeley's (2007) model of working memory. His model is divided into three components: the phonological loop, the visuo-spatial sketchpad, and the central executive. Depending upon the nature of a given mathematics problem, each of these working memory components could be involved in mathematical problem solving. However, attentional capacities reside within the central executive component and are the most "crucial feature of working memory" (Baddeley, 2007, p. 124). The central executive is "crucial" because it determines which information in a mathematics problem should receive attention and which should not (Feifer & De Fina, 2005). Baddeley (2007) and Feifer and De Fina (2005) argue that not only is the central executive critical in directing, shifting, and sustaining attention, but it is also important in the inhibition of negative distractors and the selection of necessary strategies to execute a cognitive task. The central executive, therefore, orchestrates the action required in working memory and, in turn, mathematical problem solving.

Sergeant (1996) submits that attention can be subdivided into selective attention and sustained attention. Selective attention is defined as the ability to ignore irrelevant stimuli while focusing on the task at hand (Bjorklund, 2005; Matlin, 2002; Sergeant, 1996). Sustained attention is defined as the ability to maintain one's focus as long as necessary to complete a given task (Bjorklund, 2005; Sergeant, 1996). Selective and sustained attention capacities are activated when a student begins reading a mathematics question.

For example, when reading a mathematics question, the student must discern what the question is asking while simultaneously filtering out any extraneous and irrelevant stimuli. This action is selective attention. Then, declarative and/or procedural knowledge is activated and the associated information is drawn into working memory for manipulation (Dehn, 2008; Dehn, 2010). As mathematics problems become more complex, students must maintain their focus on the question to determine the kind of problem the question is addressing, how to set up the problem, what formula(s) must be used to answer the question, and then finally how to solve the problem. This action involves both sustained attention and selective attention since the student needs to continually *select* the appropriate details to focus on while simultaneously *sustaining* their focus on the problem-solving task.

Without a combination of selective and sustained attention capacities, students are not likely to be successful problem solvers. Despite the role of attention as an initial phase in problem solving none of the cognitive diagnostic assessment models previously mentioned consider it. Given the importance of selective and sustained attention on mathematics problem solving, attention will be included as an error category in my interim assessment framework.

Once a student's attention is activated, needed information is retrieved from long-term memory and brought into working memory so that novel problem solving can occur (Dehn, 2008). Long-term memory is subdivided into episodic and semantic memory, where episodic memory consists of the relevant events in our history and semantic memory contains "all the general knowledge we possess" (Dehn, 2008, p. 72). Semantic memory is divided into declarative and procedural knowledge (Dehn, 2008).

Procedural knowledge involves knowledge about (a) how to perform actions (Dehn, 2010; Matlin, 2002), (b) how to perform the "steps required to complete various tasks" (Dehn, 2008, p. 72), (c) "how to solve problems and apply information" (as cited in Kamphaus, 2005, p. 57), and (d) the "action sequences for solving problems," or "skills, algorithms, …[and] strategies" (Rittle-Johnson & Siegler, 1998, p. 77).

Declarative knowledge includes factual information, specifically, concepts, propositions, schemata, frames, scripts, and rules (Bjorklund, 2005; Dehn, 2008; Dehn, 2010; Kamphaus, 2005; Matlin, 2002). In a mathematics context, declarative knowledge can be defined as mathematics vocabulary, mathematics facts, mathematics rules, mathematics notation and their meanings, and selection of an appropriate formula for a given problem.

According to Kamphaus (2005), declarative and procedural knowledge are inseparable components in solving problems. Dehn (2010, p. 29) argues that the "organizational structure of semantic memory lends itself well to academic learning that places a heavy emphasis on conceptual and factual learning." Thus, it is reasonable to include declarative and procedural knowledge as error categories since semantic memory errors would likely impact a student's mathematics achievement.

Attention, procedural knowledge, and declarative knowledge error categories appear to be necessary for a meaningful and cognitively diagnostic interim assessment. At this point, no other error categories will be pursued. Although each error category may be clearly defined, identifying the error category a given error exemplifies is not simple. Cognition is a complex process. Bjorklund (2005) argues that efficient information processing is an *interaction* of several processes: consciousness, attentional

capacity, and working memory. One process stimulates another in the processing of stimuli.

For instance, Kruschke (2005) posits that selective attention is involved in the categorization that is stored within declarative memory. Kruschke (2005, p. 186) used the example of deciding if an animal is a duck or a rabbit to illustrate how categorization works. Kruschke contends that when the mind is categorizing an animal as a "duck" or a "rabbit," a person needs to pay attention selectively to features of the animal to determine which characteristics it possesses and, in turn, which category it belongs to. Based upon Kruschke's theory of categorization, selective attention and declarative memory are not mutually exclusive.

Other researchers have debated the relationship between when and how conceptual and procedural knowledge develop (Rittle-Johnson & Siegler, 1998). Based on a comparison of 34 studies, Rittle-Johnson and Siegler contend that declarative and procedural knowledge do not always develop in the same manner. At times, conceptual knowledge appears to develop before procedural knowledge. At other times the converse is true. Rittle-Johnson and Siegler (1998, pp. 77-78) have postulated that conceptual and procedural knowledge "develop iteratively, with small increases in one leading to small increases in the other, which trigger new increases in the first." Given this variance in development, it is plausible that declarative and procedural knowledge are not mutually exclusive memory components. In fact, attention, declarative knowledge, and procedural knowledge are likely not mutually exclusive error categories. No research was found that indicates whether attention and procedural knowledge are mutually exclusive memory components. Therefore, one might conclude that attention and procedural knowledge are

not mutually exclusive, in as much as one has to sustain their attention on a task as one moves through solving a multi-step problem.

In brief, attention errors suggest lapses in selective and/or sustained attention abilities (Baddeley, 2007; Bjorklund, 2005; Feifer & De Fina, 2005; Matlin, 2002; Sergeant, 1996). Procedural errors are defined as calculation errors or missteps in problem solving while conceptual/declarative errors are mistakes in factual information (e.g., formulas, notations, definitions). These new error categories can potentially link students' thinking processes with their acquired skills. Thus, for any given item a teacher might remediate some students with more conceptual tasks while other students might receive remediation that is more procedurally based. This remediation process should allow for student weaknesses to be targeted according to lapses or deficits in student thinking.

Ideally, it would be helpful if these error categories could be ordered according to their degree of correctness. Since attention is more about focus and orchestration and unconcerned with the direct storage of memory components, I will assume that attention errors are lesser errors than either procedural or conceptual/declarative errors. In contrast, several studies give direction to how the ordering of the procedural and conceptual error categories could be conceptualized.

Mazzocco and Devlin's (2008) research compared students with low mathematics achievement (LA) to those with mathematical learning disabilities (MLD). Their research showed that in general students with MLD had a "weak rational number sense and inaccurate beliefs about rational numbers" whereas students with LA exhibited a partial understanding of fractions and decimals with a propensity to memorize labels,

procedures, and fraction to decimal equivalencies without a clear understanding of fundamental concepts (Mazzocco & Devlin, 2008, p. 690). This study not only gives credibility to the creation of conceptual and procedural error categories, but it also suggests that students with less mathematical skill (MLD) make more conceptual errors than students with more mathematical skill while students with slightly more skill (LA) have some conceptual and procedural understanding. Other research studies are consistent with this generalization (Geary, Hoard, & Bailey, 2011; Mazzocco, Myers, Lewis, Hanich, & Murphy, 2013). These studies suggest that conceptual errors are greater errors than procedural error.

Therefore, this study will consist of an interim assessment test design in which the items are multiple-choice, the distractors are students' common misconceptions, and the errors are aligned with an information processing theory of cognition: attention, procedural knowledge, and conceptual/declarative knowledge errors. Lastly, the errors will be ranked from most correct to least correct: attention errors, procedural knowledge errors, and conceptual/declarative knowledge errors.

Summary and Purpose

The accountability issues surrounding the NCLB Act of 2001 have prompted a relatively new phenomenon in the testing world, interim assessments. Little empirical research has occurred with interim assessments. Much of the research has relied on teacher observations, interviews, and surveys (Christman et al., 2009; Clune & White, 2008; Goertz et al., 2009; Marsh et al., 2006). Several of these studies revealed that teachers alter their instruction in response to interim assessment data (Christman et al., 2009; Clune & White, 2008), although there is substantial variability in how effective

teachers are in their data analysis and interpretation (Goertz et al., 2009). Fewer studies have investigated the effect of interim assessments on student achievement. The Carlson et al. (2011) study is one of the first large-scale empirical studies which suggest that interim assessments are a viable means of improving student achievement in mathematics. The Carlson interim assessments mirrored the state test blueprint and were administered as quarterly, predictive assessments. Training was provided to teachers and administration in data analysis, data interpretation, and the data-driven reform process. Despite the significant contributions of the aforementioned studies, no research has explored what interim assessment framework may be cognitively and instructionally meaningful for teachers. CDA models offer a possible interim assessment framework to support these goals.

Several researchers have created CDA models as a paradigm for diagnosing student content strengths and weaknesses and to potentially inform instruction (Briggs et al., 2006; Embretson, 1998; Gierl et al., 2007; Mislevy et al., 2003; Mislevy & Haertel, 2006; Rupp & Templin, 2008; Tatsuoka, 2009). The challenge is to find a testing framework that is not too fine or large grain so that the test data are not overwhelming to teachers while simultaneously being rich in diagnostic data. Six CDA models were discussed in this study: the RSM, AHM, DCM, ECD, CDS, and OMC assessments.

In brief, the RSM (Tatsuoka, 1983, 1986, 2009), the AHM (Gierl, 2007; Gierl, Leighton, & Hunka, 2007; Leighton, Gierl, & Hunka, 2004), and the DCM (Rupp & Templin, 2008; Rupp et al., 2010) are statistical models that employ Q matrix theory and attributes developed by content experts. Despite their similarities, the RSM, the AHM, and the DCM have their differences. The RSM is focused on error analysis and is often

retrofitted to tests, which compromises the cognitive diagnostic capability of the assessment. In turn, the AHM focuses on student content mastery and is applied to tests *a priori.* Although student content strengths are important, this focus may leave gaps in teachers' understanding of student weaknesses as content weaknesses are not necessarily an absence of content strengths. Even though the DCM can provide a multidimensional view of student mastery per item, the focus of this model is skill mastery. In fact, the DCM predicts the *probability* of student performance according to a profile of mastered attributes. Given that the DCM is about skills and attributes, it is not congruent with the interests of this study. The next model, ECD is a construct-centered approach which focuses on the accumulation of evidence to support student inferences (Gorin, 2007; Mislevy et al., 2003). Although the ECD provides a comprehensive approach to cognitive test development, it also appears to be a renaming of traditional test design principles and therefore, offers no new information for this study. In the next model (CDS), cognitive theory *precedes* test design and item development (Embretson, 1998, 1999, 2010; Embretson & Gorin, 2001; Gorin, 2007). Although the CDS has potential, it has been primarily used for ability measures. The last model is the OMC assessments, which given their multiple-choice format already is familiar to teachers. OMC assessment researchers recommend that distractors should be written using students' common errors and misconceptions (Briggs et al., 2006); however, this model does not include students' thinking processes in the composition of test items. In addition, the distractors do not represent error categories, but rather levels of student understanding. Although the OMC assessments are the closest test framework to the interests of this

study, changes still need to be made so that they are suitable as cognitively diagnostic interim assessments.

The purpose of this study is to create and validate an interim assessment for 7th grade mathematics. The assessment will consist of ordered multiple-choice categories with distracters that contain common student misconceptions. The error categories will be linked to cognitive processes and will comprise: attention, procedural knowledge, and declarative knowledge. Through error analysis, educators will be able to determine student's error patterns so that teachers will be better equipped to remediate their misconceptions and faulty strategies and extend their understandings. The details of the test development and its accompanying validation procedures will be described in Chapter 3.

This study will answer the following research questions:

(1) What validity evidence from the expert reviews and cognitive interviews supports the error categories?

(2) What is the relationship between students' problem-solving errors and teachers' perceptions of students' problem-solving errors?

(3) What is the item response theory evidence to support the OMC interim assessment framework?

(4) Do the errors made by advanced 6th grade mathematics students differ from those made by general 7th grade mathematics students?

(5) Do the errors made by special education mathematics students differ from those made by regular mathematics students?

Chapter III: Development and Validation of a Diagnostic Interim Assessment

Since the purpose of this study was to create and validate an interim assessment for 7th grade mathematics and to use the resulting test scores to inform instruction, validity evidence was required from multiple sources to support inferences drawn about this population (Haladyna, 2004; Haladyna & Rodriguez, 2013; Kane, 2009). As suggested by Haladyna (2004), much of this validity evidence came from a study of the item development procedures and an item analysis. Additional evidence was gathered from expert teachers' reviews and student interviews. Therefore, a mixed-methods research design was employed because quantitative and qualitative methods were necessary to answer the research questions. As stated by Creswell and Plano Clark (2011, p. 5), mixed methods research designs are appropriate when the research design "provides a better understanding of [the] research problems than either approach alone." Analyses were performed separately and then *mixed* during the discussion and interpretation of the data. Ultimately, the validation process was a joining together of the test's purpose, its inferences about the population, and the evidence gathered to support those inferences (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Kane, 2006; Schmeiser & Welch, 2006).

This chapter begins with a brief description of the research setting and a discussion of the test development process: table of specifications, item writing procedures, and error category rubric development. The second section concentrates on the types of evidence needed to support the development of the error categories,

specifically, cognitive interviews and expert teacher ratings. Following is a description of the data analysis methods needed to answer the research questions. Finally, the chapter concludes with the results of the error category validation. In contrast, chapter 4 examines the methods and results for the item analysis and the comparison of student groups (e.g., advanced 6th grade versus general 7th grade mathematics) according to the errors made. Together these chapters document the necessary evidence to support the interpretations and uses of the interim assessment scores.

Research Setting

The Virginia school division where this study occurred administered three interim assessments during the academic year. Each interim assessment was given at the end of the nine-week marking period within a 5-day testing window. Teachers elected which day within the testing window they wanted to implement the interim assessment. Immediately following interim assessment administration, teachers received score reports which described student strengths and weaknesses per the SOLs tested. The score reports also included an error analysis of student performance. Each nine week's interim assessment data informed teachers' instruction during the 2-3 days of remediation immediately following testing.

This study centered on a Virginia school division's *third nine weeks* interim assessment for 7th grade mathematics. The school division administered the 7th grade mathematics interim assessment to both advanced 6th grade and general 7th grade mathematics students. The underlying principle here was that both student groups take the same SOL test; therefore, they were given the same interim assessment. The resulting

score reports were divided by grade and class meaning there was a separate score report for advanced 6th grade versus general 7th grade mathematics students.

Test Development

The third nine weeks interim assessment consisted of 25 multiple choice items, which followed the sequencing of SOLs within the division's pacing guide. The pacing guide for this school division was detailed and specific. First, it followed the Virginia SOL curriculum framework. Second, it prescribed the sequence of content instruction and the number of days that should be devoted to each SOL. Furthermore, the pacing guide referenced division-provided instructional resources, such as manipulatives (e.g., algebra tiles, counters, fraction equivalency towers) that teachers were expected to use in their instruction. As such, division leaders supposed that all teachers would implement the pacing guide with fidelity.

In addition to items being aligned with the Virginia SOL curriculum framework and the division pacing guide, each item included three distractors that encompassed students' common errors or misconceptions. The decision to use three distractors was based on the following rationales: (a) the Virginia SOL tests use three distractors and modeling an interim assessment after the VDOE format was reasonable given that the interim assessments were meant to foreshadow student performance on the SOL test, (b) the interim assessments were meant to be used for instructional purposes, and (c) since lower-scoring examinees are typically more varied in their response patterns (Levine & Drasgow, 1983) the four- and five-option item would provide the most information (Lord, 1977). Furthermore, Lord (1977) and Levine and Drasgow (1983) posit that higher scoring examinees are less inclined to guess than lower-scoring examinees. Based on

these studies, I hypothesized that more distractors were needed to discriminate ability levels for lower-scoring test takers. Since I was interested in discerning the types of cognitive misconceptions examinees had in their problem solving in a format that paralleled the Virginia SOL exam, all items in the mathematics interim assessments had three distractors.

Table of specifications. Before test item writing commenced, a table of specifications was generated (See Table 2 and Table 3). The table of specifications indicated the quantity of items tested per SOL as well as the specific knowledge or skills each item assessed.

The interim assessments were cumulative in nature with each assessment containing 8 items from the previous interim assessment. After each interim test administration, a frequency analysis was used to determine the lowest performing items for the division. The frequency analysis measured the number and percent of students who correctly responded to each item (See Appendices A and B). The 8 items with the lowest percent correct formed the basis for the common items on the subsequent interim assessment. Because this study centered on the third interim assessment, it had 8 common items from the second nine weeks interim test. In this context, common items did not refer to the same test items, but rather *similar* test items. Common items were defined as (a) items which used the same question stem with different numbers, (b) items in which the sequence of the response options had changed so that if option A was previously correct, this was no longer the case, and (c) items which retained the same location within the overall test. I hypothesized that this common item model would be a sufficient measure of student remediation and growth in achievement.

The interim assessment addressed SOLs: theoretical and experimental probability, compound probability, statistics, relations, writing and evaluating algebraic expressions, properties of operations, arithmetic and geometric sequences, and solving linear equations. Seventeen of the 25 test items were devoted to these SOLs.

The remaining items consisted of the 8 common items as revealed by the second nine weeks frequency analysis. The advanced 6th grade mathematics frequency analysis showed that the items with the lowest percent correct included items: 11, 16, 13, 8, 25, 15, 6, and 12 (See Appendix A). In contrast, the general 7th grade mathematics frequency analysis determined that the items with the lowest percent correct were items: 11, 16, 25, 8, 6, 13, 21, and 15 (See Appendix B). These lists of lowest-performing items were nearly the same for both groups. The difference occurred with items 12 and 21. With item 21 only 31.5% of general 7th grade mathematics students correctly answered this item, while with item 12, 51% of the advanced 6th grade mathematics students answered this item correctly. Since fewer of the general 7th grade students answered item 21 correctly compared to the percent of advanced students who responded to item 12 correctly, I included item 21 as a common item. Therefore, the common items for the third nine weeks assessment followed the item structure of items: 11, 16, 25, 8, 6, 13, 21, and 15. The SOLs which corresponded to the common items included: proportional reasoning, volume and surface area of rectangular prisms and cylinders, similar quadrilaterals and triangles, properties of quadrilaterals, and transformations of polygons.

Table 2
Third Interim Assessment Table of Specifications

Number of Items	SOL	Description
2	7.2	The student will describe and represent arithmetic and geometric sequences using variable expressions.
2	7.4*	The student will solve single-step and multistep practical problems, using proportional reasoning.
2	7.5*	The student will describe volume and surface area of cylinders, solve practical problems involving the volume and surface area of rectangular prisms and cylinders, and describe how changing one measured attribute of a rectangular prism affects its volume and surface area.
2	7.6*	The student will determine whether plane figures-quadrilaterals and triangles-are similar and write proportions to express the relationships between corresponding sides of similar figures.
1	7.7*	The student will compare and contrast the following quadrilaterals based on properties: parallelogram, rectangle, square, rhombus, and trapezoid.
1	7.8*	The student given a polygon in the coordinate plane will represent transformations (reflections, dilations, rotations, and translations) by graphing in the coordinate plane.
2	7.9	The student will investigate and describe the difference between the experimental probability and theoretical probability of an event.
2	7.10	The student will determine the probability of compound events, using the Fundamental Counting Principle.
2	7.11	The student, given data in a practical situation, will construct and analyze histograms and will compare and contrast histograms with other types of graphs presenting information from the same data set.
2	7.12	The student will represent relationships with tables, graphs, rules, and words.

*Common item(s) are noted with an asterisk.

Table 3

Third Interim Assessment Table of Specifications

Number of Items	SOL	Description
3	7.13	The student will write verbal expressions as algebraic expressions and sentences as equations and vice versa; and evaluate algebraic expressions for given replacement values of the variables.
2	7.14	The student will solve one- and two-step linear equations in one variable and solve practical problems requiring the solution of one- and two-step linear equations.
2	7.16	The student will apply the following properties of operations with real numbers: commutative and associative properties of addition and multiplication, the distributive property, the additive and multiplicative identity properties, additive and multiplicative inverse properties, and the multiplicative property of zero.

*Common item(s) are noted with an asterisk.

Item-writing procedures. One middle school mathematics teacher and I created the initial interim assessment test items. Each test writer used the Virginia SOL Test Blueprint (Virginia Department of Education [VDOE], 2009a), 7th grade Mathematics Curriculum Framework (Virginia Department of Education [VDOE], 2009b), and the local 7th grade mathematics pacing guide as guides for test item content development. In addition, the test writers followed the item-writing guidelines promulgated throughout the theoretical and empirical research literature (Haladyna, 2004; Haladyna & Downing, 1989a; Haladyna & Downing, 1989b).

Haladyna and Downing reviewed 46 chapters and textbooks to derive a list of the most frequently cited item-writing rules. Then they validated the rules to determine if any should be eliminated or ameliorated. Given the scope of this study, it is noteworthy

that 19 of the studies that Haladyna and Downing (1989b) reviewed recommended that students' common errors be among the multiple choice distractors. Hence, the final list of item-writing rules for this study relied on content directives from the Virginia Department of Education and the local school division's pacing guide as well as the item-writing guidelines espoused by research.

Once the first draft of each test was developed, middle school mathematics teachers from the division reviewed the items and key through a peer review process to ensure that the items were aligned with the 7th grade mathematics SOLs and the division's pacing guide. Teachers were asked to confirm that the key was accurate and that all test items were clear, free from errors, and appropriate for 7th grade students. Revisions to items and the key were made, if necessary before test implementation.

Error category rubric development. Each item included three distractors which encompassed students' common mistakes or misconceptions. Errors were ordered according to how incorrect they were. For instance, minor errors, such as *failing to include the correct units when calculating the area or volume of a shape* would indicate more correctness than a student who was *confused by mathematics vocabulary* or one who *did not know the rule for the order of operations*. Students' scores for each item were represented as partial credit scores such that a score of 3 was completely correct, and a 2 was more correct than a score of 1. The most incorrect response was scored as 0.

In addition, the error categories corresponded to an information processing theory of cognition. Cognition, especially mathematics cognition is a complex process with no single, accepted theory describing how information is acquired, stored, and retrieved. Researchers have identified many components important to cognitive processing, such as

declarative knowledge, procedural knowledge, processing speed, and attention (Baddeley, 2007; Bjorklund, 2005; Dehn, 2008, 2010; Matlin, 2002). Processing speed was not considered as an error category since the interim assessment was an untimed test.

Based on the cognitive research previously discussed, I proposed that the error categories attention, procedural knowledge, and conceptual knowledge were *sufficient* in describing students' common mathematics mistakes. Although these three error categories were believed to be satisfactory, I did not anticipate that they described all nuances of mathematics cognition. Other categories might emerge in this study, but at this point, these three error categories appeared to be the most salient in describing students' mathematics problem solving and the source of their common errors.

Of the three error categories proposed, attention errors were the most difficult to pinpoint. In fact, I did not find any CDA literature that separated attention into its own error category. Tatsuoka (1983) focused much of the RSM on procedural errors while Briggs et al. (2006) centered the OMC assessments on progress variables which essentially captured procedural and conceptual errors. Several of the remaining CDA models (i.e., AHM, ECD, or CDS) did not explicitly address students' cognitive errors because their diagnostic capacities appeared to be more aligned with cognitive strengths.

On the other hand, Goertz et al. (2009) indicated that teachers developed four error categories in their interim assessment analysis of student performance: procedural, conceptual, other cognitive weaknesses, and contextual diagnoses. Attention errors were among the list of errors in the category "other cognitive weaknesses." The remaining errors in this category included weak reading ability, test anxiety, and low levels of English proficiency. Although each of these "other cognitive weaknesses" was very

different from each other, teachers believed they impacted mathematics achievement. Assuming this category of "other cognitive weaknesses" remained unchanged, a teacher would not know which component(s) within this category were areas the student needed strengthened. I submit that attention should be its own error category so that teachers can specifically remediate this cognitive weakness. A detailed description of how each of these cognitive error categories corresponded to mathematical problem solving follows.

***Attention errors*.** Based on the cognitive research previously discussed (Baddeley, 2007; Bjorklund, 2005; Feifer and De Fina, 2005; Matlin, 2002; Sergeant, 1996), I posited that attention errors were related to lapses in selective and/or sustained attention abilities. I suggested that attention errors were more about what a student did not do in their problem solving, rather in what they did do. Furthermore, I submitted that students with attention errors possessed both a conceptual and procedural understanding of a given mathematics problem. I proposed that students who made attention errors had (a) selected the correct formula(s), (b) understood the mathematics vocabulary and notation, and (c) correctly performed all calculations and procedures. However, somehow in their problem solving they ignored some aspect of what a question was asking them to do. Perhaps they misread the directions or they just did not follow the directions. For instance, they may have *left off* the last step when solving a problem, thinking they had completed the entire series of steps. This *leaving off of the last step* may be due to a lapse in sustaining their attention on the mathematics problem solving. In this case, the formula(s) chosen were correct, and the procedures and calculations were correct. They just did not go far enough to complete the question asked. Or, they may have been asked to simplify a fractional answer, yet they did not. Another possibility was

that they may have mislabeled the sides or angles in a geometry problem. For example, if an angle should have been labeled as $\angle ABC$, they may have transposed BC to CB and wrote the angle as $\angle ACB$. Finally, they may have been asked to find the next 2 numbers in a geometric sequence, but they only found 1 number. My hypothesis was that attention errors were more about omission rather than commission.

Procedural errors. Procedural errors were ranked here as characteristically between attention and declarative/conceptual errors. With procedural errors the student had some understanding of the underlying concepts, vocabulary, and notation; hence, they had a conceptual understanding of the mathematics concepts, vocabulary, and notation. Moreover, there was often something correct in part of the problem solving. This was why the procedural error was not seen as serious as the declarative/conceptual knowledge error. Unlike an attention error, there was a mistake somewhere in the calculation or steps performed (e.g., wrong step, skipped step). Perhaps in the calculation of a multi-step problem, the student began the problem correctly and then subsequently made a calculation error. Or, perhaps in a word problem, they understood what the problem was asking, selected the correct formula, but then they made mistakes using the formula.

Examples of procedural errors include (a) they may have calculated the slope as $\frac{\Delta x}{\Delta y}$, rather than $\frac{\Delta y}{\Delta x}$, (b) they may not have carried numbers correctly when multiplying multi-digit numbers, (c) they may not have regrouped correctly when subtracting multi-digit numbers, and (d) they may have described an ordered pair as (y, x) rather than as (x, y).

Conceptual errors. Conceptual errors were the most serious of the error categories because these errors indicated that the student had little or no understanding of fundamental mathematics concepts. Matlin (2002, p. 254) described declarative or conceptual knowledge as "knowledge about facts and things." In a mathematics context conceptual errors could entail selecting the wrong formula, misunderstanding mathematics vocabulary or mathematics notation. Many times the errors referred to the student's absence of some factual information. For example, suppose a student was asked to graph a series of points in a Cartesian coordinate plane. If the student reversed the x- and y-coordinates for *all* of the points, it could be deduced that the student did not know which axis was the x-axis or y-axis. On the other hand, if the student did not reverse all of the coordinates, it was possible that there was some other type of error. Perhaps, in this case, the student was confused about graphing only points on the axes, not all points. Graphing some of the points correctly and others incorrectly could be a procedural error since the student demonstrated an understanding for graphing some points correctly. Table 4 lists a brief description of each error category and its corresponding partial credit assignment.

Table 4
Partial Credit Scores and Cognitively Diagnostic Error Categories

Partial Credit Score	Error Category Descriptor	Brief Synopsis of Error Categories
3	Correct	No error
2	Attention errors	This error is more about what the student did not do, rather than what they did do. The student may have misread the directions or they did not follow the directions. All of the mathematics performed is correct.
1	Procedural knowledge errors	Parts of the problem solving are correct. Errors include calculation errors or procedural mistakes—wrong step, skipped step. This student has a conceptual understanding of the mathematics required in an item. They select the correct formula and are familiar with the needed mathematics vocabulary and notation.
0	Declarative or Conceptual knowledge errors	Factual knowledge errors. Errors may involve conceptual understanding, mathematics facts, vocabulary, or the meaning of notation. These are concepts that can be memorized and later recalled—factual knowledge. They also may not be able to recognize the type of mathematics problem a given item represents. They may choose the wrong formula.

Error category assignment. As discussed in Chapter 2, the error categories were not believed to be mutually exclusive (Kruschke, 2005; Rittle-Johnson & Siegler, 1998). The overlap and influence of one error category on another could make the assignment of error categories arduous. I suggest when categorizing each incorrect response, the designated error category must be the *primary* or overarching cognitive error. In this way, the most prominent cognitive feature of the error should be recognized. Correct error category assignment is very important to the eventual success of teachers' remediation efforts. Incorrect error category assignment could result in a teachers' misunderstanding of student mistakes and the selection of inappropriate remediation interventions.

Cognitive Interview and Expert Teacher Reviews

In order to confirm the development of the error categories and their corresponding assignment for each test item's distractors, a series of cognitive interviews and expert teacher reviews were established.

Cognitive interviews. Cognitive interviews, or more commonly called "think-alouds," comprise two forms: concurrent versus retrospective interviews (Ericsson & Simon, 1993). In the context of this research study, concurrent think-alouds (TA) are cognitive interviews which occur *while* a student is solving mathematics problems. With a retrospective TA the student explains their thinking about solving mathematics problems *after* a test is administered.

At the beginning of a cognitive interview, the researcher typically has the student practice how to "think aloud" while solving several mathematics problems until they are more comfortable with the process (Ericsson & Simon, 1993). Researchers often provide directions such as "keep talking" if a student becomes silent (Ericsson & Simon, 1993). Probes may be used by the researcher to encourage richer, more detailed speech from the student. Ericsson and Simon (1993) contend that the "thinking aloud" process is not foreign to students since they often are expected to explain their thinking in a classroom setting. However, concurrent and retrospective TAs are not without their challenges.

For instance, although with a concurrent TA the student explains their problem solving as they calculate mathematics problems, the vocalization of thoughts may slow their problem solving (Ericsson & Simon, 1993). On the other hand, the retrospective TAs give the student the opportunity to problem solve without the interview demands interrupting their cognitive processing. However, if retrospective TAs occur too long

after a test is administered, some of the episodic memories associated with the problem solving may deteriorate. To capture the benefits of both TA methods, Ericsson and Simon (1993) recommend including concurrent and retrospective TAs in a research study. I employed retrospective TAs for my cognitive interviews because I believed they would be the least invasive to the learning environment and their data is generally consistent with those rendered with concurrent TAs (Ericsson & Simon, 1993).

Student think-alouds. A student think-aloud protocol was administered to advanced 6th grade and general 7th grade mathematics students. The goal of the TA was (a) to deepen my understanding of student's mathematics misconceptions and (b) to confirm the development of the interim assessment's error categories. Since students were recalling past problem solving, probes were required. Probes are often used to clarify student's statements or as a means of re-enacting the problem-solving episode. In this study, probes constituted questions such as, "tell me why you moved from this step to this step?" or "can you show me what you were thinking?" or "why did you not select answer A?" Answers to some of these probes helped illuminate a student's problem-solving rationale and gave direction to why students struggle mathematically. The resulting student data was used to confirm the development and assignment of the error categories.

Expert teacher reviews. Expert teacher reviews occurred in two forms: teacher think-alouds and teacher ratings of the error category assignments. Teacher TAs were structured in the same way as the student TAs except they addressed teacher perceptions of students' cognitive processing. In other words, what did the teachers believe students

were thinking mathematically when they chose certain item distractors? A complete description of both types of expert teacher reviews follows.

Teacher think-alouds. The teacher think-aloud protocol was administered to three middle school mathematics teachers. The objective of the TA was to ascertain what teachers perceive is the reasoning behind student problem-solving errors. Probes were used to clarify teacher's statements or to deepen their explanations of student problem-solving behaviors. Probes consisted of questions such as, "what do you think is the reason a student might choose answer A?" This data was helpful in determining how closely teacher perceptions of student errors fit with the error categories. Additionally, this data allowed the researcher to compare educators' preconceptions of student errors with students' explanations of their reasoning.

Teacher raters. Three middle school mathematics teachers were trained according to the error category rubric. The training comprised a theoretical explanation of each error category as it related to cognitive processing. Several test examples were provided of incorrect responses and their error category assignment. The teacher raters were given the opportunity to practice categorizing errors on several items before the rating data was collected. I fielded questions to clarify the error category rubric, if necessary. Then, the teachers rated all 25 test items using the interim assessment items and the answer key *without* error categories. Since each item had three incorrect responses, each teacher rated 75 incorrect response options. The resulting teacher rating data was analyzed to determine the interrater agreement.

Participants. The student participants for the cognitive interviews (i.e., student think-alouds) consisted of nine advanced 6th grade and general 7th grade mathematics

students. Three mathematics teachers participated in the teacher think-alouds and three different mathematics teachers participated as teacher raters. The teachers had at least three years of teaching experience. As such, they had an in-depth understanding of common student misconceptions and middle school mathematics.

Data collection procedures. Parental consent forms were distributed to all advanced 6th grade and general 7th grade mathematics students at three middle schools prior to interim assessment administration. Furthermore, teacher interview and rater consent letters were distributed to all selected mathematics teachers prior to the interim assessment administration. Before testing commenced, all consent forms were collected by the researcher. A stratified random sample of students was selected from those forms in which parents bestowed consent. Stratification occurred in two ways: middle school matriculation and course enrollment (i.e, advanced 6th grade mathematics and general 7th grade mathematics). This double stratification ensured that students from each course were represented in a given school's random sample. Since three students were selected from each school, a total of nine students were selected for the TAs. Student assent procedures were followed to ensure the willingness of each student to participate in the cognitive interview process. Student assent consisted of written assent via the parent consent letter and verbal assent the day of the cognitive interview.

Within approximately one week after the interim testing window commenced, the researcher traveled to the three middle schools to perform the student TA interviews. Since the mathematics teachers chose which day within the testing window they administered the interim assessment, the number of days from when the students actually took the interim assessment varied.

The sampled TA students were removed from the mathematics classroom and brought to a secure testing location where they were able to freely explain their problem solving within the TA interview. These students were asked to explain their reasoning for one third of the test, or approximately eight test items. Student 1 at each school was interviewed for items 1-8, student 2 at each school was interviewed for items 9-16, and student 3 was interviewed for items 17-25.

A student think-aloud protocol was utilized giving each interview a consistent structure. Each think-aloud was audio-taped and transcribed to ensure the interview methodology was followed. The think-aloud interview and the general testing for all students occurred within a 50-minute class period. At the conclusion of the TA, each student received a $20 gift card for their participation in the interview process.

Following the interim assessment administration, three advanced 6th grade and general 7th grade mathematics teachers were interviewed using a think-aloud protocol. As with the student TAs, the teacher TAs were audio-taped and transcribed to ensure the interview methodology was followed. Educators examined the same test items the sampled students were given for their TAs. Each teacher received a $20 set of mathematics manipulatives for their participation.

After the student and teacher TAs were completed, the three teacher raters were trained. Once the teacher raters were comfortable with how to rate the distractors, they were given them a copy of the interim assessment questions and a copy of the key *without* the error category assignments so they could assign error categories to each distractor. Teacher raters were permitted to use the aforementioned training resources to better assist them in assigning error categories to each distractor.

Error Category Validation Results

Results in this section address the research questions: *what validity evidence from the expert reviews and cognitive interviews supports the error categories* and *what is the relationship between students' problem-solving errors and teachers' perceptions of students' problem-solving errors*. This section begins with an examination of the student participants and proceeds with an analysis of the student cognitive interviews, or think-alouds. Following is an examination of the teacher interview participants and a chi-square analysis of their perceptions concerning students' problem-solving behaviors. Subsequently, teacher rater demographics and interrater agreements are presented.

Cognitive interviews. Three students from each of three middle schools participated in retrospective cognitive interviews (think-alouds). In other words, 9 student think-alouds were administered. Of these 9 students, 4 were black students (44.4%), 4 were white students (44.4%), and 1 was an Asian student (11.1%). Two students (22.2%) received special education accommodations. Table 5 displays the student interviewee demographics.

Table 5
Demographics of Student Interviewees

Student	Gender	Ethnicity	Special Education	Course Enrollment
1	F	White		6th Advanced
2	F	Black		6th Advanced
3	F	White		6th Advanced
4	M	White		7th Regular
5	F	Asian		6th Advanced
6	F	Black		7th Regular
7	F	Black	x	7th Regular
8	F	Black	x	7th Regular
9	F	White		7th Regular

In the cognitive interviews students were asked to not only describe what they were thinking when they selected an incorrect response, but they were also asked why they did *not* select each of the remaining response options. Transcriptions of the student think-alouds were analyzed for patterns and trends using case-ordered matrices (Miles & Huberman, 1994). Descriptive statistics were calculated to characterize the quality of student mathematics talk. If students' mathematics talk fit neatly with one of the error categories, then this error category was assigned to the student's explanation. If not, they were assigned an "other" category. Two types of types of explanations fell within the "other" category: process of elimination and guessing (See Table 6). Combined all students provided 225 responses to the distractors, yet only 6 responses (2.7%) were due to a process of elimination and 7 responses (3.1%) were because of guessing.

Table 6
Descriptive Statistics of Student Mathematics Talk

Data	Proportion of Student Responses	Percent
Confusion with "NOT" (*items 8 and 23)*	2/6	33%
Misunderstanding "comparing" (*item11)*	2/3	67%
"Other" category		
__process of elimination	6/225	2.7%
__ guessed on 1-2 items	7/225	3.1%

Evidence of the process of elimination and guessing can be seen in the dialogue which follows.

Process of Elimination:

Researcher:	Why did you pick B?
Student:	Um. Well, I knew it wasn't either A or D just by looking at that.
Researcher:	By looking at what?
Student:	At the um, -2x – 3 = 7
Researcher:	So, you immediately eliminated A and D because they didn't match the -2x?

Student: Yeah.

Researcher: When you say they didn't match, can you point to where they didn't match?

Student: Um, well the -2x would be the shaded –x. So, there's two.

Researcher: Okay. And A didn't have that and neither did D.

Student: Yeah.

Researcher: Ok. So, you're pointing to the variable tiles [that they] were not the same size or shading.

Guessing:

Student: Well, I did analyze everything but I was really confused and I was trying, and I kept second guessing myself and everything. So, I was like, well BC/DE isn't really comparing anything so if it's not comparing anything, then they're equal. So, I kind of just put that one in desperation.

Other than assigning error categories to student responses, student responses were analyzed to determine if there were any other patterns or trends. Two patterns emerged which were related to specific words. Several students struggled with the words *not* and *comparing*. Although items 8 and 23 both used the word *not* in the question stem, only item 8 proved to be a challenge. Despite the fact that in both cases the word *not* was in all capital letters, one student never saw it as evidenced in the following discourse.

Researcher: So, why did you put B?

Student: I think that's wrong, is it wrong?

Researcher: It's wrong.

Student: I thought so. I didn't see the NOT.

Researcher: Ahhhh. That's important.

In contrast, another student misunderstood the meaning of *not* in item 8.

Researcher: Can you tell me why you put A?

Student: I put A because it said it could not be…It said the proportions could not be used to find out how many students got a B. So, I put A, because it is not a proportion to equal anything. So, you

	don't know how many students got how many of 6 out of 8 students got a B on the report card.
Researcher:	Okay, so do you know what A is called, if it is not a proportion?
Student:	No, it's just a fraction.
Researcher:	Is there another name for that?
Student:	Um, no.

Item 11 used the word *comparing* in the question stem, which confused two students. The question asked students to compare squares and rectangles and then select the true statement. This student thought comparing two things meant they could not be alike, yet *comparing* means to find the similarities and dissimilarities of two or more things.

Student:	B, I didn't really touch.
Researcher:	Why is that?
Student:	Both a square and a rectangle are a special type of rhombus. That has nothing to do between the differences.
Researcher:	Okay, so now we're on C, which is what you picked.
Student:	Uh huh. Squares have 4 equal sides. Rectangles have 2 pairs of equal sides. I thought that this was right, because it's true, but I got it wrong because it wasn't comparing.
Researcher:	What do you mean?
Student:	It's not, um, squares have all 90 degree angles, but they have 4 equal sides. Rectangles have 90 degree angles, but they don't have all equal sides. That's like comparing, but what I chose was just a fact, which wasn't right.

Next, I compared the items students answered incorrectly with the error categories assigned on the test key. Collectively, student participants missed 43 test items. Overall, the explanations students provided regarding their errors matched 74.4% of the error categories assigned on the test key (See Table 7).

Table 7: Agreement between Student Explanations and the Test Key

Student's Error Explanations	Test Key Error Category Assignments			
	Conceptual	Procedural	Attention	Frequency
Conceptual	13	1	6	20
Procedural	0	16	1	17
Attention	1	2	3	6
Total	14	19	10	43

Specifically, if a conceptual error was assigned on the test key, student explanations agreed with the key 92.9% of the time, otherwise student explanations agreed with an attention error (7.1%). If a procedural error was assigned on the test key, student agreement was 84.2%, otherwise student explanations agreed with an attention error (10.5%) or a conceptual error (5.3%). But, if an attention error was assigned on the test key, student explanations agreed with the test key 30% of the time, otherwise student explanations agreed with a conceptual error (60%) or a procedural error (10%).

A chi-square analysis was performed to determine the relationship between incorrect student responses and the error categories assigned on the test key. Henceforth, student's incorrect item responses will be referred to as a *student mistake*. For example, if an item's correct answer is C and a student answered B, then B is a *student mistake*. The contingency table in Table 8 compares *student mistakes* with the error category assignments on the test key. The chi-square analysis revealed that there was no statistically significant difference in students' descriptions of their mistakes and the error categories assigned on the key ($\chi^2 = 3.309,\ df = 2,\ p = 0.191$). In other words, the explanations students provided regarding why they made mistakes fit with the error categories assigned on the test key.

Table 8
Student Mistakes vs Error Categories Assigned on the Test Key

	Conceptual	Procedural	Attention	Total
Student Mistakes	20 (0.7)	17 (-0.1)	6 (-0.9)	43
Test Key: Error Assignment	14 (-0.7)	18 (0.1)	11 (0.9)	43
Total	34	35	17	86

Note: Standardized residuals in parentheses

In the student interviews I questioned students about their mistakes *and* their reasons for not selecting the remaining response options. However, I discovered that fewer students were able to adequately articulate their reasons for *not selecting* a distractor. As such, some students' responses consisted of threads of mumbling, disorganized, or nonsensical speech. These responses illustrated how unorganized some students are in their mathematical thinking. Thus, for this next analysis I only included those students who could adequately explain their thinking. Five out of the nine students interviewed were selected for this analysis.

The explanations students provided of *why they did not select a specific distractor* agreed with the error categories assigned on the test key 86.8% of the time. If a conceptual error was assigned, agreement between the student's explanation and the key was 95.8%. If a procedural error was assigned, agreement was 92.3%. But, when an attention error was assigned, agreement dropped to 63%. A contingency table compared student explanations of each distractor they did not select to the error categories assigned on the test key (See Table 9). A chi-square analysis revealed that there was no statistically significant difference between student explanations of distractors they did not select and the error categories assigned on the test key ($\chi^2 = 1.671,\ df = 2,\ p = 0.434$).

Table 9
Articulate Student Explanations vs Error Category Assignment

	Conceptual	Procedural	Attention	Frequency
Articulate Student Explanations	56 (0.6)	38 (-0.1)	20 (-0.7)	114
Test Key: Error Category Assignment	48 (-0.6)	39 (0.1)	27 (0.7)	114
Total	104	77	47	228

Note: Standardized residuals in parentheses

Expert teacher reviews. Expert teacher reviews were comprised of (a) teacher perceptions of students' problem-solving behaviors and (b) teacher ratings of the error category assignments. Each of the next two sections begins with a description of the teacher participants followed by a descriptive analysis of the data.

Teacher think-alouds. Three middle school mathematics teachers participated in the teacher cognitive interviews, or think-alouds. Teachers were selected who had many years of experience teaching advanced 6th grade and/or general 7th grade mathematics. Teacher demographics are depicted in Table 10.

Table 10
Demographics of Teacher Interviewees

Teacher	Gender	Total Yrs Teaching	Yrs Teaching 6th or 7th grade Mathematics
Teacher 1	F	30	23
Teacher 2	F	7	7
Teacher 3	F	23	23

Teacher transcripts were organized in case-ordered matrices. Transcripts were reviewed multiple times to discover the patterns and trends in teacher perceptions. Like the student responses, teacher responses were coded according to the error category

assignments. Table 11 compares teacher perceptions with the error category assignments on the test key.

Table 11
Teacher Perceptions vs Error Category Assignments on the Test Key

	Key-Conceptual	Procedural	Attention	Frequency
Teachers-Conceptual	90	0	0	90
Procedural	1	78	1	80
Attention	0	6	45	51
Total	91	84	46	221

Overall, teacher perceptions of student problem-solving behaviors agreed with the error category assignments on the test key 96.38% of the time. If a distractor was scored as a conceptual error, teacher agreement was 98.9%. If a distractor was scored as a procedural error, teacher agreement was 92.86%. With distractors scored as attention errors, teacher agreement was 97.83%.

A contingency table compared teacher perceptions of students' problem-solving behaviors to the error category assignments (See Table 12). A chi-square analysis showed that there was no statistically significant difference between the teacher perceptions and the error category assignments ($\chi^2 = 0.100,\ df = 2,\ p = 0.951$).

Table 12
Teacher Perceptions vs Error Categories Assignments on the Test Key

	Conceptual	Procedural	Attention	Frequency
Error Categories	31 (0.1)	28 (0.1)	16 (-0.2)	75
Teacher Perceptions	90 (0.0)	80 (-0.1)	51 (0.1)	221
Total	121	108	67	296

Teacher perceptions were then compared to *student mistakes* using a contingency table (See Table 13). Overall, agreement between teacher perceptions and student mistakes was 74.2%. If a student made a conceptual error, teacher perceptions agreed

with the students 92.86% of the time. If a student made a procedural error, teacher agreement was 86.54%. However, if a student made an attention error, teacher agreement was 32.35%.

A chi-square analysis was performed to determine the relationship between *student mistakes* and teacher perceptions of their problem-solving behaviors. The chi-square analysis indicated there was a statistically significant difference in students' descriptions of their mistakes and teacher perceptions of students' problem solving ($\chi^2 = 8.139\ df = 2,\ p = 0.017$). No standardized residuals were statistically significant. The effect size suggested a weak association between student mistakes and teacher perceptions (Cramer's V =0.178).

Table 13
Student Mistakes vs Teacher Perceptions

	Conceptual	Procedural	Attention	Frequency
Student Mistakes	60 (1.3)	50 (-0.1)	18 (-1.6)	128
Teacher Perceptions	42 (-1.3)	52 (0.1)	34 (1.6)	128
Total	102	102	52	256

Note: Standardized Residuals in parentheses

Next, teacher perceptions were compared to *students' explanations of all item distractors.* As mentioned previously, I only included those students who could adequately explain their thinking. Overall, when students explained each item distractor they agreed with teacher perceptions 84.95% of the time. For instance, if teachers perceived an error was conceptual, student responses agreed with them 81.82% of the time. If a teacher perceived an error was procedural or attention, student agreement was 87.39% and 87.5%, respectively.

A contingency table compared articulate student explanations of the distractors to teacher perceptions (See Table 14). A chi-square analysis revealed that there was a

statistically significant difference between students' distractor explanations and teacher perceptions ($\chi^2 = 6.450,\ df = 2,\ p = 0.040$). No standardized residuals were statistically significant. In addition, the effect size was small indicating a weak association between student distractor explanations and teacher perceptions (Cramer's V = 0.104).

Table 14
Articulate Student Explanations of Distractors vs Teacher Perceptions

	Conceptual	Procedural	Attention	Frequency
Articulate Student Explanations	132 (1.0)	111 (0.2)	56 (-1.5)	299
Teacher Perceptions	111 (-1.0)	107 (-0.2)	81 (1.5)	299
Total	243	218	137	598

Note: Standardized Residuals in parentheses

Teacher raters. Three *different* middle school mathematics teachers rated the interim assessment's error category assignments. Teacher rater selection was based on two factors: the teacher had previous experiences in item development for the school division's interim assessments and the teacher had several years' experience teaching advanced 6th grade and/or general 7th grade mathematics (See Table 15). Although rater 3 had one year of *mathematics* teaching experience, she served as my mathematics specialist intern where she was exposed to test development and data analysis methods.

Table 15
Demographics of Teacher Raters

Teacher	Gender	Years Teaching	Years Teaching Mathematics	Years Teaching 6thAdvanced or 7thRegular	Other Related Mathematics Experiences
Rater 1	M	6	6	6	n/a
Rater 2	F	16	16	4	n/a
Rater 3	F	4	1	1	Graduate intern

All three teacher raters had perfect agreement with me for 24 of the 25 test items, resulting in a raw agreement of 100%. Item 7 was the exception. Here raters 2 and 3 were in perfect agreement with me, while rater 1 was in perfect agreement with me except for 1 distractor (See Table 16). Rater 1 rated this distractor as a procedural error whereas I rated it as a conceptual error. Consequently, for item 7 the agreement between the 3 raters and me was 88.9%.

Table 16
Item 7 Error Category Assignments

Me	Rater 1			Total
	0	1	2	
0	1	1	0	2
1	0	0	0	0
2	0	0	1	1
Total	1	1	1	3

Chapter IV: Test Scoring, Analysis, and Group Comparisons

Whereas Chapter 3 looked at the validation of the error categories and evidence to support the test development process, Chapter 4 explores the methods and results concerning an item analysis and a comparison of student groups (e.g., advanced 6th grade and general 7th grade mathematics). And so, this chapter begins with an examination of the test participants and data collection procedures. Then, the methods and results of the item analysis are discussed, specifically, classical test theory (CTT), distractor analysis, differential item functioning (DIF), and item response theory (IRT) methods. Following is a look at the methods and results for the student group comparisons. The goal of the student group comparisons is to provide direction to teachers' instructional planning and remediation efforts by illuminating potential differences among student groups in the errors they make.

Participants

The participants were advanced 6th and general 7th grade mathematics students and their teachers from an urban school division in Virginia. Although these students were in different grades, they each were taught 7th grade SOLs and their teachers followed the same curriculum pacing guide schedule throughout the academic year. Thus, advanced 6th grade mathematics and general 7th grade mathematics students were administered the same interim assessments. By sampling students from both of these mathematics courses, a more heterogeneous population of students could be achieved.

Approximately, 547 students were enrolled in advanced 6th grade and general 7th grade mathematics courses (Advanced 6th = 259 and Regular 7th =288). This population

of students was characterized by (a) a high level of poverty, about 60% for the division, and (b) a low VA SOL pass rate on the new mathematics SOLs, less than 25% for 7th grade students during the 2011-2012 academic year (Virginia Department of Education website, n.d.). The VA state average SOL pass rate for 2011-2012 was 58% (Virginia Department of Education website, n.d.). Due to the transiency of this population, student course enrollment numbers were apt to vary by the time the interim assessments were administered. Thirteen licensed educators taught these courses (Advanced 6th grade = 5 teachers, General 7th grade mathematics = 8 teachers).

Data collection procedures

All of the division's advanced 6th grade and general 7th grade mathematics students' third interim assessment scores were collected after the third nine weeks interim assessment administration. Since data were collected as part of the normal educational process in the division, parent consent was not required. Once the score data were collected, all of the data files were organized for the item analysis. An item analysis of the interim assessment scores was performed using the psychometric analysis program jMetrik (Meyer, 2002). Since the test items were polytomous where "partial credit [was awarded] for partial success," the IRT partial credit model was employed (Masters, 1982, p. 150). An important contribution of the IRT analysis was an item map (Wilson, 2005). Here the item map served two functions: as a means of validating the ordering of the error categories and to give teachers an additional diagnostic tool to assist in their remediation efforts. Furthermore, the error categories were also validated using a chi-square analysis, which compared the IRT theta values to the error category assignments on the test key.

Item Analysis Results

The results presented here represent the quantitative evidence that support the development of a mathematics interim assessment with cognitively diagnostic error categories. Results in this section address the research question: *what is the item response theory evidence to support the OMC assessment framework?* Since this section provides the technical documentation of the interim assessment, it begins with an examination of the student participants. Following is a complete item analysis of the test scores to include a classical test theory analysis, distractor analysis, differential function analysis, and item response theory analysis.

Participants. The 3rd 9 weeks interim assessment was administered to 519 advanced 6th grade and general 7th grade mathematics students. Of these 519 students, one student's data was removed because all of the student's responses could not be read by the Scantron scanner. As a result only 518 student responses were retained for data analysis.

The student population was comprised of 255 males (49.2%) and 263 females (50.8%). Students were distributed among several race categories: 4 American Indians (0.8%), 12 Asians (2.3%), 252 Blacks (48.6%), 212 Whites (40.9%), and 38 Multi-Race (7.3%) students. Student enrollment was somewhat greater in general 7th grade mathematics than in advanced 6th grade mathematics, 281 students (54.2%) and 237 students (45.8%), respectively. 47 out of 518 students (9.1%) received special education accommodations.

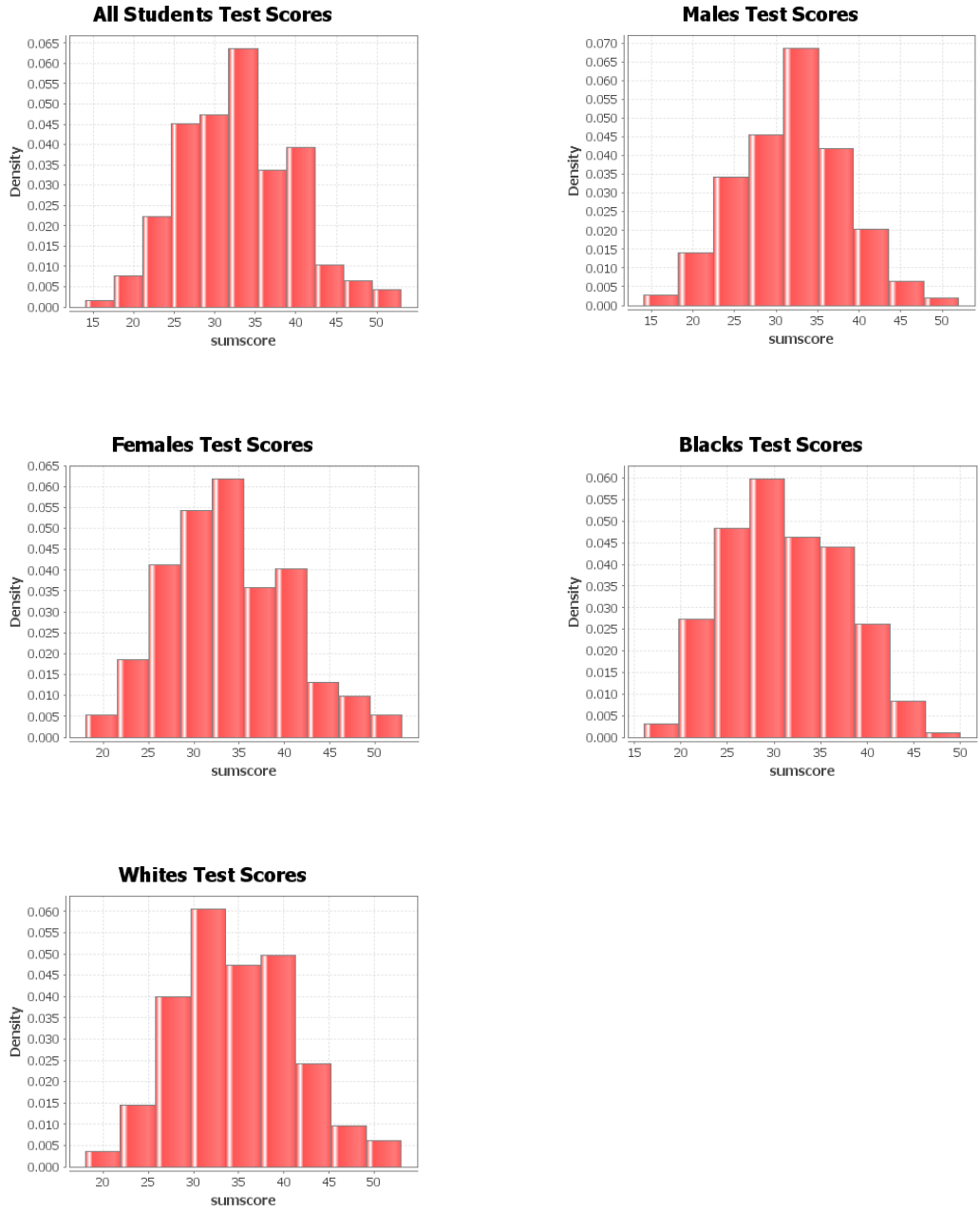
As mentioned previously, test items were polytomously scored based on the cognitive error categories. Correct items were scored as 3, attention errors were scored as 2, procedural errors as 1, and conceptual errors as 0 (See Appendix C). Because items were written using students' common errors, three error categories were not necessarily conducive to all items. As such, 8 items involved conceptual, procedural, and attention error categories, while 15 items had only two error categories. Because items 5 and 23 focused on student understanding of mathematics vocabulary, conceptual errors were the only relevant errors assigned. Hence, items 5 and 23 were dichotomous. With 25 items and correct responses scored as 3 and the partial credit scoring rubric seen in Appendix C, the maximum possible sum score was 56. The groups *all students, males, females, blacks,* and *whites* had similar mean test scores and standard deviations (See Table 17).

Table 17
Test Score Descriptive Statistics

Group	N	Minimum Score	Maximum Score	Mean	Standard Deviation
All Students	518	14	53	32.8	6.80
Males	255	14	52	32.1	6.75
Females	263	18	53	33.6	6.77
Blacks	252	16	50	31.1	6.26
Whites	212	18	53	34.8	6.71

A comparison between each subgroup's test scores is more easily visible in a histogram. The distribution of each group's test scores is unimodal and symmetric (See Figure 4).

Figure 4: Histogram of All Group's Test Scores



Classical Test Theory Analysis. An item analysis was performed using jMetrik (Meyer, 2002), which revealed that item difficulties ranged from 0.3166 to 2.5541. The easiest item was item 2 (difficulty= 2.5541) and the most difficult item was item 23

(difficulty= 0.3166) (See Appendix D). Livingston (2006) suggests that if test items are partial credit items, it is helpful to report the frequency of students who responded to each item distractor (See Appendix E).

Item discriminations ranged from -0.0174 to 0.4079. Item 16 was the only item with a negative discrimination, -0.0174. Consideration may be given to removing item 16, which would increase reliability to 0.6459. Four items had acceptable discriminations between 0.30 and 0.70 (Allen & Yen, 1979; Bond & Fox, 2007). Thirteen test items had low discriminations between 0.20 and 0.29 and four items had discriminations between 0.10 and 0.19. Together the low discrimination items lowered the estimated reliability because they did not discriminate well between students who did and did not possess the assessed skills (Allen & Yen, 1979; Livingston, 2006). Three items had very low item discriminations (< 0.10), which meant that these items discriminated very poorly between ability groups. These items included item 6 (0.0914), item 9 (0.0934), and item 23 (0.0521). All of the items with low item discriminations were reviewed to determine if their item construction might be ameliorated.

The interim assessment was assumed to be unidimensional, although this assumption was not specifically tested. Reliability estimates are in excess of 0.62 for all subgroups, except black students (See Table 18). The greatest score variation was seen with black students (SEM= 4.19 and 4.25), while white students saw the smallest score variation (SEM= 4.04 and 4.09). In addition, the 95% confidence intervals indicate that the reliability coefficients rendered comparable results across *most* subgroups. For instance, Guttman's lambda-2 estimates were between 0.5495 and 0.6352.

Table 18
Guttman's Lambda-2 Reliability Estimates and SEM

Group	Coefficient	95% CI	SEM
All Students	0.6251	(0.5769, 0.6702)	4.1572
Male	0.6255	(0.5559, 0. 6888)	4.1237
Female	0.6277	(0.5596, 0.6897)	4.1250
Blacks	0.5495	(0.4653, 0.6261)	4.1913
Whites	0.6352	(0.5605, 0.7025)	4.0416

Distractor analysis. The premise of distractor analysis is to improve distractors and overall item performance by determining which distractors should be refined or rewritten and which distractors should simply be eliminated. This, in turn, provides "important validity evidence to the overall validation process" (Haladyna, 2004, p. 219). The distractor analysis was conducted using statistical, tabular, and graphical data analysis methods. Even though each of these methods is uniquely different in their approach, they should provide similar evidence of problematic item distractors.

Distractor-total correlations. A distractor-total correlation should be negative. However, the distractor analysis revealed that items 8, 9, and 11 each had one distractor with a positive distractor-total correlation (See Table 19). These items warranted further review.

Table 19
Distractor Analysis of Item Discriminations

Item #	Item discrimination	A	B	C	D
8	0.1751	-0.2862	-0.3524	**0.2997**	0.0066
9	0.0934	**0.0760**	0.0882	-0.4096	-0.1904
11	0.2517	-0.3980	-0.3874	0.0477	**0.1975**

Note: Item discrimination for the correct response is in bold.

Only one similarity could be found among the three positive item discriminations and item construction. For items 9 and 11, the distractors with positive item

discriminations were those assigned as attention errors. Perhaps because these items dealt with attention to detail, more students who were not strong in the assessed domain were enticed by these distractors. Otherwise, the content domain assessed for items 8 (proportional reasoning), 9 (probability), and 11 (geometry) were quite different. Furthermore, with item 8 the positive item discrimination occurred with the distractor endorsed the *least* frequently while with item 11 it was the distractor endorsed the *most* frequently. Conversely, item 9's positive item discrimination was neither the most nor least frequently endorsed. Apart from the similarity with two attention distractors having positive item discriminations, there appears to be no other discernable pattern with item construction.

Student response rate. Haladyna (1993) posits that item distractors with less than a 5% response rate are likely selected due to guessing and, therefore, should be rewritten. However, before immediately rewriting these items, a more thorough analysis needs to occur to determine why there is less than a 5% response rate. Items for this assessment in which the distractors have < 5% response rate are seen in Table 20.

Table 20
Distractors with < 5% Response Rate

Item #	Item Difficulty	A (%)	B (%)	C (%)	D (%)
2	2.5541	4.8	**75.3**	9.2	10.4
3	2.3649	58.2	1.0	1.2	**39.5**
4	1.3745	3.9	**66.7**	11.8	17.3
5	0.6931	**69.2**	16.2	12.5	1.5
7	0.6988	**33.9**	54.3	9.4	1.9
22	2.3938	**66.5**	10.4	18.7	4.2

Note: Correct responses are in bold.

Several of the items which rendered distractors with less than a 5% response rate were among the easier test items, specifically, items 2, 3, and 22. It is reasonable that

easier items would not be enticing to students because most students would likely know the required content. Distractors with a low response rate tended to be conceptual errors (items 2, 3, 5, and 22), although attention errors (items 4 and 7) and procedural errors (item 3) were also evident. The content domains assessed included: items 2, 3, and 4 (algebra), items 5 and 7 (properties of numbers), and item 22 (probability). Finally, with items 5, 7, and 22 the distractor with the low response rate was option D. In reviewing the content of these distractors, these are the distractors that were the least plausible.

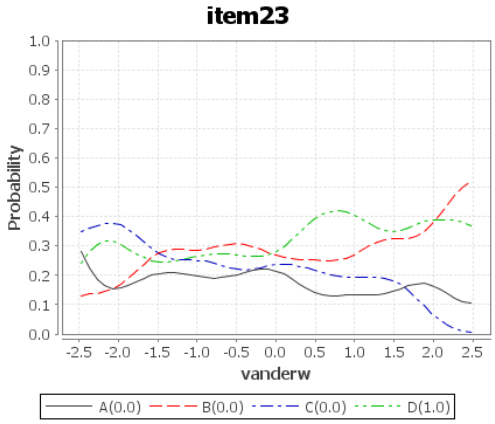
Nonparametric curves. Unlike parametric curves, nonparametric curves are not required to conform to strictly monotonically increasing functions. As such, nonparametric curves can reveal more about problematic items (Haladyna, 2004, 2013; Meyer, in press) than parametric curves. Each item was examined to see if they exhibited one or more of the following undesirable traits: (a) correct response curves that decrease as student ability increases (Haladyna, 2004; Meyer, in press), (b) correct responses that are never endorsed, and (c) distractors that are flat or non-discriminating (Haladyna, 2013).

To create nonparametric curves, sum scores were transformed into van der Waerden normalized scores. The nonparametric curves revealed that generally all test items displayed the highest score category increasing monotonically as student ability increased. There were slight deviations in this pattern with item 23 (See Figure 5).

Item 23 was an application of statistics vocabulary. Students were asked to determine which type of graph could *not* be used for a given histogram. Item 23 had a very gradual overall increase across ability with peaks at ability scores of approximately

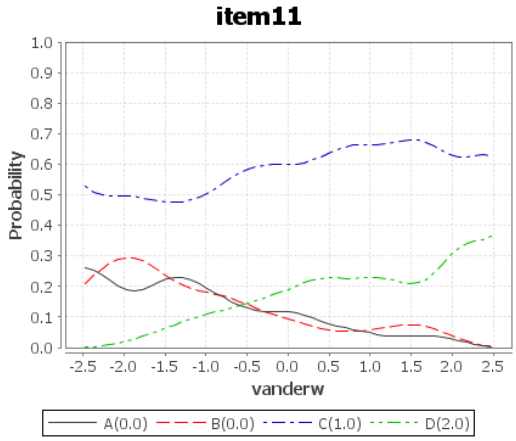
-2.25 and 0.75. All response options were more likely at some point except option A which was "*line plot*."

Figure 5: Correct Response Option Decreases Slightly as Ability Increases



In addition, the highest score category was more likely in all items except item 11 (See Figure 6). With item 11 students were expected to select the true statement comparing squares and rectangles. Distractor C was chosen far more frequently across all ability groups.

Figure 6: Correct Response Option is Never More Likely

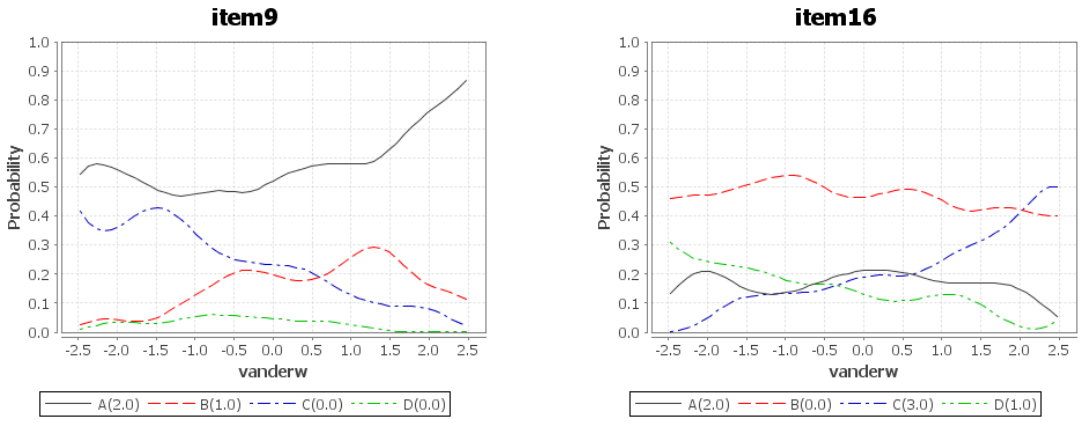


Distractor C was "*Squares have 4 equal sides. Rectangles have 2 pairs of equal sides."* If a student did not read the distractor carefully, they might have missed that the word *opposite* was not included in the sentence. To be correct the statement should have said, *"Squares have 4 equal sides. Rectangles have 2 pairs of opposite equal sides."* Rather, it said *"2 pairs of equal sides."* Two pairs of equal sides could refer to *adjacent sides* being equal; hence, the error. This distractor was enticing to 59.3% of students (See Appendix E).

All of the test items had at least one response option that was never more likely. Six items had nearly flat distractor curves (items 3, 5, 7, 9, 19, 25) (See Appendix F). Flat curves suggest that a distractor is poorly functioning and needs to be rewritten since no ability level is endorsing this distractor with any frequency. In fact, each of these distractors has less than a 5% response rate, as previously discussed. With 2/3 of these items, the flat curve occurred with response option D. Six items had one item distractor that was non-discriminating (items 1, 6, 8, 14, 15, 23) (See Appendix G). With 2/3 of these items, the non-discriminating curve occurred with a procedural error. In all but one of these cases, there were 2 procedural response options and 1 option that was either a conceptual or attention error.

Items 9 and 16 were especially problematic items (See Figure 7). Item 9's correct response was *always* the more likely response while item 16's correct response was only the most likely response with greater than a 2.0 van der Waerden score. With item 9 students were asked to calculate the probability of 2 dependent events. Given that 54.2% of the students answered this item correctly, it was likely not too difficult for many students.

Figure 7: Correct Response Options for Items 9 and 16



Differential item functioning analysis. Differential item functioning (DIF) analyses were performed to measure the degree to which item performance is the same for members of two different groups that have the same overall ability level. jMetrik (Meyer, 2002) was used to perform the DIF analyses. jMetrik uses the Cochran-Mantel-Haenszel statistic for testing statistical significance and the standardized mean difference to assess practical significance among polytomous items (Meyer, in press). Items were classified using the ETS classification system where AA items indicate negligible DIF, BB items suggest moderate DIF, and CC items signal severe DIF.

The first DIF analysis involved male and female examinees (focal group = females, reference group = males). The gender DIF analysis revealed that no items were classified as CC items. On the other hand, item 25 displayed moderate DIF (BB+) and favored females. Given that item 25 is needed for content validity and it displays a moderate degree of DIF, it was retained.

The second DIF analysis measured test fairness with respect to race (focal group = blacks, reference group = whites). The race DIF analysis indicated that item 17

exhibited severe DIF (CC-) and favored whites. Furthermore, although 10 items exhibited moderate DIF (BB+ or BB-), the DIF is *canceling DIF* (See Table 21).

Table 21
Differential Function Analysis for Gender and Race

DIF Analysis	Item #	Mantel-Haenszel	p-value	Effect Size	DIF Classification
Gender	25	1.31	0.25	0.12	BB+
Race	4	4.58	0.03	0.19	BB+
Race	7	1.88	0.17	0.15	BB+
Race	11	4.45	0.03	-0.18	BB-
Race	14	0.22	0.64	-0.10	BB-
Race	16	2.43	0.12	0.23	BB+
Race	17	5.26	0.02	-0.20	CC-
Race	18	2.78	0.10	0.25	BB+
Race	21	1.41	0.23	-0.27	BB-
Race	22	6.24	0.01	-0.19	BB-
Race	24	3.13	0.08	-0.11	BB-
Race	25	2.54	0.11	0.18	BB+

Canceling DIF describes situations where there are an equal number of BB+ and BB- items resulting in no overall DIF. Because item 17 exhibited severe DIF, a purified sum score was calculated. Then, the DIF analysis was run a second time with the purified matching score, but the results were worse. Here item 11 was classified as CC- and several items were classified as BB+ or BB- items without the benefit of canceling DIF. Next, a nonparametric curve for item 17 was created to determine the uniformity in the DIF (See Figure 8). Nonparametric curves displayed uniform DIF across all ability levels except at -2.5 where the reference and focal curves merged briefly. Given the severity of DIF and that other items within the test assess the same content domain item 17 could be eliminated from the test.

Figure 8: Race DIF for Item 17

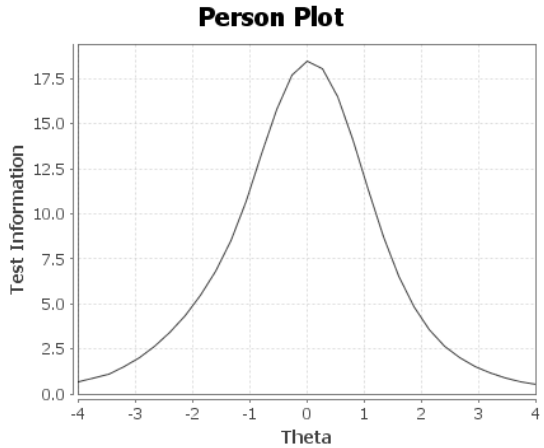
Item response theory analysis. Each item was scored using a partial credit model (PCM) and cognitively diagnostic error categories. As previously discussed, a correct response was scored as 3, an attention error was scored as 2, a procedural error was scored as 1, and a conceptual error was scored as 0. All 25 test items did *not* incorporate all 3 error categories. Because each distractor was written using common errors, all 3 error categories were not relevant or "common." Furthermore, if items did not use all three error categories, not only was the error scoring not necessarily sequential, but scoring did not always include 0. Therefore, items *without* 3 error categories needed to be recoded (See Appendix C).

The PCM results revealed that estimated item difficulties ranged from approximately -1.04 to 1.03 with a standard error of approximately 0.05 (See Appendix H). The weighted mean square (WMS) or infit statistics were from 0.87 to 1.26 and the unweighted mean square (UMS) or outfit statistics were from .8051 to 1.3875. Since the interim assessment was a low-stakes assessment, good infit (WMS) and outfit (UMS) statistics should be between 0.7 and 1.3 (Bond & Fox, 2007, p. 243). Infit and outfit

measures for item fit were all within acceptable limits except for item 16 (See Appendix H). Although item 16's infit was reasonable at 1.26, the outfit statistic = 1.39 indicating that item 16 underfits the model (Bond & Fox, 2007, p. 240). In other words, this item was likely too difficult for students resulting in their erratic responses. The root mean square of the error (RMSE) = .06 and indicates a reasonable, although not ideal model fit for this population (Browne & Cudeck, 1992). Item reliability was estimated at 0.9863 with a separation index = 8.4971. Thus, item difficulty estimates are very reproduceable with this sample and the measure can sufficiently separate items according to their item difficulty.

Person ability (theta values) estimates ranged from -0.8881 to 2.2549 with a mean theta = 0.2709 and a standard deviation = 0.4276. Person reliability was estimated at 0.6597 with a person separation index = 1.3922. Person reliability was likely low due to the homogeneity of the population assessed and the use of common errors. Person separation was low given that it is less than 2 (Meyer, in press). The test information function indicated that test precision was highest at a theta = 0 (See Figure 9).

Figure 9: Test Information Function



Next, the ordering of the error categories was validated through an item map and an examination of the error category thresholds (See Appendix I). Items in Appendix I were grouped according to the error categories assigned. For example, all items scored exclusively with *conceptual and procedural errors, procedural and attention errors,* or *conceptual and attention errors* are in separate tables. Only 6 items did not have a reversal: items 6, 8, 11, 17, 19, and 24. A few patterns were evident in items without reversals versus those with reversals. First, items without reversals tended not to have a "throwaway" or implausible distractor. Second, nonparametric curves suggested that *generally* items without reversals tended to have steeper decreasing slopes for the distractors (See Appendix J). Finally, 2/3 of the items without reversals were scored with conceptual and procedural errors. Only one item was scored with procedural and attention errors while one item was scored with conceptual and attention errors. No other patterns were apparent with respect to content domains assessed or item difficulty. Several researchers have suggested that there is an ordering to conceptual and procedural errors (Geary et al., 2011; Mazzocco & Devlin, 2008; Mazzocco et al., 2013). Perhaps the reversals were somehow related to the attention errors.

An item map visually depicted the ordering of the error categories (See Appendix K). Here several test items had threshold reversals. Higher point values mapped to low points on the scale and lower point values mapped to higher points on the scale. Out of 23 polytomous test items, 16 items had attention errors as one of their error categories. These items included: 1, 2, 3, 4, 7, 9, 10, 11, 13, 14, 16, 17, 18, 20, 21, and 22. Only 12.5% of the aforementioned items did not have a reversal, items 11 and 17. In contrast, 7 items *did not* have attention errors as one of their error categories. These items

consisted of: 6, 8, 12, 15, 19, 24, and 25. Approximately, 57.1% of these items did not have a reversal, items 6, 8, 19, and 24. On the other hand, 8 items had all three error categories: items 1, 2, 3, 13, 16, 18, 21, and 22. All of these items have reversals as demonstrated in the item map. With 62.5% of these items, the attention error was the reason for the reversal. These items included: 1, 2, 13, 18, and 22.

The item map also revealed that many of the test items were closely matched to student ability. Because of the relationship between persons, item difficulty, and the step parameters, these students were measured with a great amount of precision. Conversely, several students at the highest ability levels and some students at the lowest levels were measured with less precision and more error. In order to measure these students with greater precision additional items need to be added to the test at the higher and lower ends of the ability scale.

Student Group Comparisons According to Types of Errors

In the current accountability climate, teachers are required to use test data to drive instruction, but many teachers are in a quandary how to achieve this expectation (Mandinach & Honey, 2008; Marsh et al., 2006; Young, 2006). This requirement can be exacerbated if teachers teach more than one course preparation (e.g., general 7th grade mathematics and Algebra 1) or more than one level of the same course (e.g., advanced versus general 7th grade mathematics). Coupled with the differences in course preparations are the demands to differentiate instruction for low achieving students, special education students, second-language learners, and general education students. Anecdotally, these teachers may make comparisons between some of the aforementioned student groups (e.g., general versus special education students, or advanced versus

general mathematics students), such as *what are the cognitive strengths of the various groups? what are their shared cognitive weaknesses?, should I use different instructional strategies for one group versus another to increase achievement?* If, for instance, teachers had evidence that special education students make more conceptual errors as a group than general education students, this data could be helpful in teachers' instructional planning and remediation. In addition, comparing student errors according to ability levels could provide further evidence of an ordering to the error categories. It is questions like these that suggest an investigation needs to be made comparing the student groups in this study.

The student participants in this study can be divided into several student groups: (a) advanced 6th grade versus general 7th grade mathematics students, (b) general education versus special education mathematics students, and (c) student ability quartiles. I am interested in discerning if any of these student groups differ in the types of errors made. If so, this data could provide teachers with valuable lesson planning/remediation data. For example, if general 7th grade mathematics students as a group made more conceptual errors than advanced 6th grade mathematics students, this data would suggest that teachers place greater emphasis on conceptual knowledge constructs (e.g., mathematics vocabulary, mathematics facts, mathematics notation, and general mathematics concepts) with the general 7th grade mathematics students. Therefore, I compared student groups (e.g., advanced 6th grade versus general 7th grade, and special education versus general education) to determine if there were significant relationships between group membership and the types of errors made on the 7th grade interim assessment.

Analysis of Student Groups. Student data in this study were divided into several groups to determine if student membership affected the kinds of errors students made. These groups were (a) advanced 6th grade and general 7th grade mathematics, (b) special education and general education mathematics students, and (c) student ability quartiles. Chi-square analyses were performed using SPSS (IBM SPSS Version 22.0, 2013).

Student Group Comparison Results

Results in this section address the research questions *do the errors made by advanced 6th grade mathematics students differ from those made by general 7th grade mathematics students* and *do the errors made by special education mathematics students differ from those made by general mathematics students?* This section begins by examining the relationship between course enrollment and the volume of student errors. Following is an investigation which explores the frequency of errors made by special needs students and general education students. The last section explores the association between student ability and errors made.

Advanced 6th Grade versus General 7th grade Mathematics Students. Student test scores were divided among two groups, those enrolled in advanced 6th grade mathematics and those enrolled in general 7th grade mathematics. A contingency table was created to compare course enrollment and the frequency of conceptual, procedural, and attention errors for all items (See Table 22). A chi-square analysis revealed that the there was a significant association between the types of errors made and course enrollment ($\chi^2 = 6.838,\ df = 2,\ p = 0.033$). No standardized residuals were significant. Furthermore, the effect size was small indicating that the relationship between course enrollment and types of errors made was very weak (Cramer's V = 0.032).

Table 22
Course Enrollment vs Error Categories Assigned on the Test Key

	Conceptual	Procedural	Attention	Frequency
Adv 6th grade	1115 (-0.5)	981 (-0.9)	786 (1.7)	2882
Gen 7th grade	1567 (0.4)	1416 (0.8)	969 (-1.4)	3952
Total	2682	2397	1755	6834

Note: Standardized Residuals in parentheses

Special education versus general education mathematics students. Student test scores were divided into two groups, those with and without special education accommodations. Next, a contingency table was created to compare the frequency of errors made by special needs and general education students (See Table 23). A chi-square analysis revealed that there was not a significant association between the types of errors made and whether or not a student has special education accommodations ($\chi^2 = 0.222,\ df = 2,\ p = 0.895$). In addition, Cramer's V = 0.006, which indicates there is no relationship between special/general education and the types of errors made.

Table 23
Special Education/General education vs Error Categories Assigned on the Test Key

	Conceptual	Procedural	Attention	Frequency
Special Education	264 (-0.1)	236 (-0.2)	175 (0.4)	675
General education	2395 (0.0)	2161 (0.1)	1528 (-0.1)	6084
Total	2659	2397	1703	6759

Note: Standardized Residuals in parentheses

The results from this analysis were surprising given that across many school divisions special needs students as a group tend to have a lower pass rate on the Virginia SOL than other subgroups (Virginia Department of Education website, n.d.). However, special needs students by definition encompass not only students with low ability levels, but also those with physical and emotional handicaps. Therefore, it might be more appropriate to divide student test scores into quartiles according to IRT theta scores. This

would allow for a chi-square analysis comparing ability with the error category assignments. This analysis would also provide additional validity evidence with respect to the ordering of the error categories.

Ability quartiles versus error categories. Student test scores were divided into quartiles with respect to students' IRT theta scores and compared to the error categories for all items (See Table 24). A chi-square analysis indicated that there was a significant association between ability quartiles and errors made ($\chi^2 = 84.340,\ df = 6,\ p < .001$). Several standardized residuals were statistically significant. These residuals suggest that: Q1 students tend to make more conceptual errors than higher-level students ($z = 2.4$, $p < 0.05$) and Q4 students tend to make more attention errors than lower-level students ($z = 5.6$, $p < 0.001$). Despite the significant standardized residuals and chi-square, the association between ability quartiles and errors made was very weak (Cramer's V = 0.079).

Table 24
Theta Ability Quartiles vs Errors Made for All Items

	Conceptual	Procedural	Attention	Frequency
Q1	824 (2.4)	724 (1.8)	384 (-5.0)	1932
Q2	669 (1.4)	541 (-1.1)	405 (-0.5)	1615
Q3	773 (-1.6)	752 (0.8)	559 (1.0)	2084
Q4	416 (-2.5)	377 (-2.1)	407 (5.6)	1200
Total	2682	2394	1755	6831

Note: Standardized Residuals in parentheses

Although there are significant differences in ability and the kind of errors made, how much of that difference is attributed to whether an item has all 3 error categories, or just two? Consequently, I grouped items that shared the same error category assignments. Then, I performed a chi-square analysis on each group of items comparing ability quartiles and errors made.

The first contingency table compared ability quartiles and all items that had conceptual, procedural, and attention errors (See Table 25). A chi-square analysis showed that there was a significant association between ability quartiles and errors made when the errors encompassed three error categories ($\chi^2 = 46.172$ $df = 6$, $p < .001$). Several standardized residuals were statistically significant. These residuals suggested that Q1 students tend to make more conceptual errors than higher-level students (z = 3.2, p < 0.01). Moreover, Q3 students (z = 2.5, p < 0.5) and Q4 students (z = 2.4, p < .05) tended to make more attention errors than lower-level students. In spite of the significant standardized residuals and chi-square statistics, the association between ability and the conceptual, procedural, and attention errors was very weak (Cramer's V = 0.107).

Table 25
Ability Quartiles vs Items with Conceptual, Procedural, and Attention Errors

	Conceptual	Procedural	Attention	Frequency
Q1	219 (3.2)	216 (1.0)	180 (-3.7)	615
Q2	136 (-0.3)	170 (0.8)	180 (-0.5)	486
Q3	142 (-1.8)	177 (-1.0)	260 (2.5)	579
Q4	77 (-1.7)	96 (-1.0)	152 (2.4)	325
Total	574	659	772	2005

Note: Standardized Residuals in parentheses

The second contingency table compared ability quartiles and items that incorporated conceptual and procedural errors (See Table 26). Here a chi-square analysis revealed that there was a significant association between ability and the conceptual and procedural errors made ($\chi^2 = 14.842$, $df = 3$, $p = .002$). Some standardized residuals were statistically significant. These residuals suggested that Q2 students tended to make more conceptual errors than higher-level students (z = 2.0, p < 0.05) while Q2 students made fewer procedural errors than other students (z = -2.0, p < 0.05). Although the chi-

square and residuals were significant, Cramer's V suggested a weak association between ability and the occurrence of conceptual and procedural errors (Cramer's V = 0.088).

Table 26
Ability Quartiles vs Items with Conceptual and Procedural Errors

	Conceptual	Procedural	Frequency
Q1	286 (0.5)	252 (-0.5)	538
Q2	264 (2.0)	188 (-2.0)	452
Q3	291 (-1.6)	328 (1.7)	619
Q4	159 (-0.6)	164 (0.7)	323
Total	1000	932	1932

Note: Standardized Residuals in parentheses

The third contingency table compared ability quartiles with items including procedural and attention errors (See Table 27). A chi-square analysis showed that there was a statistically significant association between ability quartiles and procedural and attention errors ($\chi^2 = 28.488,\ df = 3,\ p < 0.001$). Several standardized residuals were also statistically significant. In fact, Q1 students tended to make more procedural errors than higher-level students ($z = 2.2$, $p < 0.05$) while Q4 students were more likely to make attention errors than their lower-level cohorts ($z = 3.1$, $p < 0.01$). Nevertheless, the association between ability and procedural and attention errors was weak (Cramer's V = 0.143).

Table 27
Ability Quartiles vs Items with Procedural and Attention Errors

	Procedural	Attention	Frequency
Q1	256 (2.2)	129 (-2.6)	385
Q2	183 (-0.2)	138 (0.2)	321
Q3	250 (0.1)	178 (-0.2)	428
Q4	117 (-2.6)	141 (3.1)	258
Total	806	586	1392

Note: Standardized Residuals in parentheses

The final contingency table investigated the association between ability quartiles and items with conceptual and attention errors (See Table 28). A chi-square test revealed

that there was a significant association between ability quartiles and errors made ($\chi^2 = 39.735,\ df = 3,\ p < .001$). Several standardized residuals were statistically significant. For instance, Q1 students tended to make more conceptual errors than higher-level students ($z = 2.1$, $p < 0.05$) whereas Q4 students tended to make more attention errors than lower-level students ($z = 4.1$, $p < 0.001$). Despite the statistical significance of these measures, there was a weak association between student ability and errors made (Cramer's $V = 0.196$).

Table 28
Ability Quartiles vs Items with Conceptual and Attention Errors

	Conceptual	Attention	Frequency
Q1	190 (2.1)	75 (-2.7)	265
Q2	162 (0.7)	87 (-0.9)	249
Q3	195 (0.0)	121 (0.0)	316
Q4	89 (-3.2)	114 (4.1)	203
Total	636	397	1033

Note: Standardized Residuals in parentheses

Chapter V: Discussion

This last chapter begins with an overall summary followed by interpretations and conclusions of the validity evidence. Subsequently, three limitations are explored and suggestions offered for further research.

Summary

The effects of NCLB (No Child Left Behind Act of 2001, 2002) and the reauthorization of the ESEA of 1965 (US Department of Education, 2010) have created a data-driven instructional environment within public schools as evidenced by the school improvement literature (Mandinach & Honey, 2008; Marsh et al., 2006). Teachers and administrators feel the strain of not only increasing students' mathematics achievement but doing so while also increasing the rigor of their instruction as they prepare students for state EOG/EOC tests. Should schools and school divisions not meet state achievement targets, they are threatened with school sanctions and takeovers (Hu, 2011; Rundquist, 2013). This accountability climate has pushed teachers to demand that all required assessments be of "maximum instructional value" (Huff & Goodman, 2007, p. 24). Not surprisingly, many school divisions have turned to interim assessments as a means of meeting their achievement targets despite the paucity of research to support this venture (Goertz et al., 2009). The attractiveness of interim assessments lies in the promise of providing teachers with instructionally useful data about student achievement *prior* to students sitting for their state EOG/EOC test. Cognitively diagnostic interim assessments offer a potential framework to support the needs and demands of educators within this accountability climate.

Cognitive diagnostic assessments (CDAs) are assessments of student learning which diagnose student "knowledge structures and cognitive processing skills" so that remediation is informed (Leighton & Gierl, 2007b, p. 3; Nichols, 1994). Although several CDA paradigms exist, such as the RSM (Tatsuoka, 1983), AHM (Gierl et al., 2007), DCM (Rupp & Templin, 2008), ECD (Mislevy et al., 2003), CDS (Embretson, 1998), and OMC (Briggs et al., 2006), the challenge is to find a testing framework that is rich in diagnostic data while not overwhelming teachers with too much data.

In order to provide teachers with an interim assessment that offered "maximum instructional value," I developed an interim assessment framework which partly followed the OMC model, but it also was influenced by the interim assessment, school improvement, and cognitive psychology literature (Baddeley, 2007; Goertz et al., 2009; Mazzocco & Devlin, 2008; Mazzocco et al., 2013). For instance, all items were multiple-choice with four response options. Distractors incorporated students' common errors and were cognitively diagnostic. Error categories were tied to conceptual knowledge, procedural knowledge, and attention constructs. The goal of the interim assessment framework was to use error analysis to provide teachers with cognitively rich data with which they could improve student achievement in mathematics.

Therefore, the purpose of this study was to create and validate an interim assessment for 7th grade mathematics and to use the resulting test scores to inform instruction. As recommended in the research literature, validity evidence was gathered from multiple sources to support diagnostic inferences about the population (Haladyna, 2004; Haladyna & Rodriguez, 2013; Kane, 2009). Evidence came from an item analysis,

distractor analysis, cognitive interviews, and expert teacher reviews. This mixed-methods study sought to answer the following research questions:

(1) What validity evidence from the expert reviews and cognitive interviews supports the error categories?

(2) What is the relationship between students' problem-solving errors and teachers' perceptions of students' problem-solving errors?

(3) What is the item response theory evidence to support the OMC interim assessment framework?

(4) Do the errors made by advanced 6^{th} grade mathematics students differ from those made by 7^{th} grade mathematics students?

(5) Do the errors made by special education mathematics students differ from those made by general mathematics students?

Conclusions

The conclusions rendered here represent a synthesis of all of the validity evidence gathered in this study. Conclusions are presented in three broad areas: validation of the error categories, item analysis, and implications for mathematics instruction.

Validation of the Error Categories. This section addresses the research questions: *what validity evidence from the expert reviews and cognitive interviews support the error categories* and *what is the relationship between students' problem-solving errors and teachers' perceptions of students' problem-solving errors*. Student error explanations were *generally* congruent with the error categories assigned on the test key. For example, whether students were explaining their *mistakes* (i.e., incorrect item responses) or *why they did not select specific distractors*, student explanations were about

90% in agreement with the conceptual and procedural error categories. If, however, students were explaining a mistake and an attention error was assigned, agreement was only 30%. Conversely, if students were explaining *why they did not select a specific distractor* and an attention error was assigned, agreement was 63%. Clearly, most of the variation in error category assignment was with *attention errors*. When a chi-square analysis was performed it revealed that there was no significant difference between student explanations of their mistakes and the error categories assigned on the test key. Thus, overall the error categories were validated by the students' cognitive interviews.

Teachers participated in cognitive think-alouds where they provided their perceptions of *why students selected a given item distractor*. When teacher's perceptions were compared to the error category assignments on the test key, teachers agreed with the error category assignments about 96% of the time. Yet, when teacher's perceptions were compared against student's *mistakes* or *student's explanations of distractors*, agreement dropped to about 74% over the three error categories. Chi-square analyses revealed that there was no significant difference between teacher perceptions and the errors on the test key whereas there was a significant difference when teacher perceptions were compared to student's actual errors. Again, the greatest variation was with the *attention errors*. Finally, teacher raters were given a test and a test key with no error categories. After receiving training in the error categories, teacher raters assigned error categories to each distractor. The agreement between the teacher raters and the test key was nearly 100%. Clearly, the common thread throughout the student and teacher cognitive interviews is the *attention errors*. More research needs to be done with respect to refining the definition of the attention error category.

Taken together, there is substantial evidence validating the conceptual and procedural error categories. Although there was some validity evidence for the attention error category, the question becomes *is the attention error category too narrow (or too broad)?*

Item analysis. This section focuses on the research question: *what is the item response theory evidence to support the OMC interim assessment framework*. An item analysis was performed using classical test theory (CTT), distractor analysis, differential item functioning (DIF) analysis, and item response theory (IRT) methods.

Although some analyses suggested that items 6, 9, and 23 might be problematic. This was not generally confirmed across all of the analyses. Item 17 was shown to have uniform DIF favoring white students over black students. No other analyses indicated that item 17 was an issue, yet this result is significant and should not be ignored. Consideration should be given to removing this item or rewriting it.

On the contrary, several analyses revealed that item 16 was problematic. The CTT analysis showed that item 16 had a negative item discrimination which indicated that lower-scoring students had a greater chance of scoring the item correct than higher-scoring students. This could be due to the test item being misskeyed or ambiguously written. The test key revealed that the item was not misskeyed. Moreover, item 16 underfit the IRT partial credit model. Underfitting items are a concern because it can indicate that the item was too difficult for the population resulting in students' erratic responses (Bond & Fox, 2007, p. 240). A nonparametric curve of item 16 revealed only students at the highest ability level were more likely to respond correctly to this item. Based upon this item's construction and the various statistical and graphical analyses,

item 16 should be rewritten. Currently, item 16 uses 2 graphics, a photo of a garage and a floor plan. The item asked students to find the number of gallons required to paint the interior of the garage: the walls, floor, and ceiling. This problem was about surface area, yet many students calculated the volume. Perhaps a labeled, 3-D drawing would help make this item more clear for students.

Nonparametric curves revealed that generally all test items displayed the highest score category increasing as student ability increased. The highest score category was the most likely response in all items except item 11 where an attention error was the most likely response. Nonparametric curves likewise showed that all of the test items had at least one response option that was never more likely. This could be due to the use of students' common errors. Several items had a flat curve or non-discriminating curve, which indicated that the distractor was not performing well and should be rewritten or eliminated. Many times the flat curves occurred with option D. Perhaps some students never read option D, because the item was easy or they were "enticed" by a different response option. Or, option D was so implausible to some students that it was rarely selected. Moreover, the flat curve could be due to the use of common student errors. Changing the sequence of the response options *could* change the functionality of the distractor, but it is more likely that these response options need to be rewritten.

A distractor analysis of response rates indicated that 6 items had less than a 5% response rate. Haladyna (2004) argues this is likely due to guessing. Option D was the response option with less than a 5% response rate for three of the six items. Option D was also the least plausible distractor in each of these three items. Altogether the nonparametric curves and the distractor analysis results suggested that 2 distractors may

be sufficient to measure the latent trait and still provide a cognitive diagnosis of student ability. Thinning the number of distractors fits with Haladyna's (1993) research.

Reliability estimates were generally lower than expected using classical test theory methods (i.e., Guttman's lambda2 = 0.6251) and item response theory methods ($\rho = 0.6597$). Several factors likely contributed to the low reliability estimates.

First, most of the test items using CTT methods had item discriminations below 0.30, which is important given that item discriminations between 0.3 and 0.7 maximize test score reliability (Allen & Yen, 1979). Second, Sadler (1998) argues that reliability and item discriminations are decreased when *common errors* are used as distractors. Sadler contends that common error distractors include attractive answers which entice students uncertain of the correct response. As such, these items are more difficult. Third, the population assessed comprises a Virginia urban school division in which less than 25% of 7th grade mathematics students passed the SOL test in 2012. Overall, this population is *generally* low performing and homogeneous. The current 7th grade population is likely similar in many respects to previous 7th grade students. Haladyna (2004) suggests that item discriminations can be biased when the population assessed is homogeneous.

Validity evidence to support the ordering of the error categories was sought from two sources: an item map and a chi-square analysis comparing student ability and the error categories. An item map revealed that several items had reversals. The most important finding in the item map was that of the 8 items which have conceptual, procedural, and attention errors, 5 of them have the threshold between procedural and attention errors reversed. This suggests that the attention error category was not

consistently the least serious error among the error categories. Furthermore, a chi-square analysis revealed that there was a statistically significant difference between student ability and the error category assignments. Although the association between these variables was weak, lower-level students tended to make more conceptual errors than higher-level students and higher-level students tended to make more attention errors than lower-level students. The conceptual error results are supported by the extant literature (Geary et al., 2011; Mazzocco & Devlin, 2008; Mazzocco et al., 2013).

Even though there is some evidence to support the ordering of the error categories (i.e., conceptual errors are generally more serious than procedural ones), the attention error category evidence is not consistent enough. At times the attention error category is the least serious error. At other times it is the most serious error. More research needs occur to determine if indeed there is an ordering to the error categories, especially with respect to the attention error.

Implications for mathematics instruction. This section answers the research questions: *do the errors made by advanced 6th grade mathematics students differ from those made by general 7th grade mathematics students* and *do the errors made by special education mathematics students differ from those made by general mathematics students*. First, the chi-square analysis comparing errors made by advanced 6th grade and general 7th grade mathematics students revealed that there was a significant difference in the errors made. General 7th grade mathematics students tended to make more conceptual errors than advanced 6th grade mathematics students and advanced 6th grade students tended to make more attention errors than their general 7th grade cohorts. Not surprisingly, the greatest inconsistency in the data was with *attention errors*.

Second, there were no significant differences in errors made between special education and general education mathematics students. This is likely because special education students encompass not only those with low ability levels, but also those with emotional and physical handicaps.

Finally, there were significant differences in errors made by students at the lowest ability quartile, quartile 1. These students tended to make more conceptual errors than their higher-ability cohorts. In turn, the highest-ability students tended to make more attention errors than the lowest-ability students.

Taken together these results suggest that teachers spend more time teaching and/or remediating conceptual knowledge constructs (e.g., mathematics vocabulary, notation, and facts) to lower-ability students, especially those enrolled in lower level mathematics courses. These generalizations fit with the extant literature (Geary et al., 2011; Mazzocco & Devlin, 2008; Mazzocco et al., 2013).

Limitations

One of my early research design decisions was to implement concurrent and retrospective cognitive interviews as espoused by Ericsson and Simon (1993). Cognitive interviews were to occur during the third nine weeks interim assessment administration, but access could not be achieved at that time. In fact, access was not achieved until shortly after the third nine weeks interim assessment administration resulting in solely retrospective cognitive interviews. The danger with retrospective cognitive interviews is that students might forget how they solved specific test items. Forgetting can be mitigated by having artifacts that stimulate someone's memory. For instance, during the cognitive think-alouds, I had access to the students' actual paper-pencil test during the

interview. When students saw their test this frequently helped trigger their memory of how they solved specific test items. Sometimes I needed to ask a student to re-enact "answering an item" during the interview, because they could not remember how they solved it by simply looking at their test. Usually, this helped alleviate their forgetting. On a rare occasion, a few students could not remember how they solved a test item despite my probes or requests. I, then, proceeded to ask them about the remaining distractors and why they think they did not select them.

A second limitation was that access to each student was limited to a 50-minute class period. Immediately before class began I retrieved the student interviewee and brought them to the interview location. The entire interview, including practice time needed to be completed within 50 minutes per the agreement with the school division. This was especially challenging with students whose processing speed was slow or whose thinking was unorganized and unsystematic.

A third limitation was that despite the school division's directive that all mathematics teachers follow the pacing guide and use the manipulative tools and other resources provided, some teachers likely deviated somewhat from those expectations. Instructional fidelity is not easy to achieve given that each teacher directs the instructional environment within their classroom. In other words, teachers decide the dosage of content delivered as well as the instructional strategies, manipulatives, and methods employed. Since I did not have access to teacher identification variables, I was not able to discern the extent to which teacher's instruction and "dosage" contributed to this data set. Including teacher identification variables would be an appropriate next step for a future study.

Opportunities for Further Research

There are two directions in which future research needs to move. First, more validity research needs to occur with respect to the error categories: conceptual, procedural, and attention. Although this study did provide evidence that helped validate the conceptual and procedural error categories, the attention error category needs further research. For instance, cognitive think-alouds *did* reveal that students do have lapses in their attention while solving mathematics problems. But, this was not always the case. At other times, some students had difficulty articulating their mathematical problem-solving. Their thinking was disorganized and sometimes nonsensical. What is it about some students' thinking that they have difficulty relating their thoughts? It appears that more is going on here than an attention error can appropriately define. Perhaps the attention error category is really more about weaknesses with respect to executive function. Executive function as defined by Naglieri and Goldstein (2013) includes nine components: *attention*, emotion regulation, flexibility, inhibitory control, initiation, organization, planning, self-monitoring, and working memory. Not only is attention subsumed within executive function but so are planning, organization, self-monitoring, and working memory. Presumably, all of these components contribute toward the act of mathematical problem solving as well as the explanation of it. If indeed the attention error category was too narrow, this might explain the lack of consistent agreement with the attention error category between educators and students. Thus, a future study could incorporate concurrent cognitive think-alouds which are directed at validating the error categories: conceptual, procedural, and *executive function*.

Second, teachers not only want required assessments to provide "maximum instructional value," they also want instructional strategies to accompany student diagnostic data (Huff & Goodman, 2007, p. 24). Thus, new research needs to explore interventions tied to each of the error categories. In this way teachers will have appropriate guidelines and tools to help redirect students' common errors and misconceptions and extend student understandings. Without unique instructional strategies linked to the error categories, teachers will likely remediate students' errors using item-by-item teaching (Shepard, 2010) and "procedural" teaching strategies (Goertz et al., 2009; Oláh et al., 2010).

References

Ainsworth, L., & Viegut, D. (2006). Common formative assessments: How to connect standards-based instruction and assessment. Thousand Oaks, CA: Corwin Press.

Allen, M. J., & Yen, W. M. (1979). Principles of test construction. In *Introduction to measurement theory* (pp. 118-147). Long Grove, IL: Waveland Press.

Alonzo, A. C., & Steedle, J. T. (2009). Developing and assessing a force and motion learning progression. *Science Education*, *93*(3), 389-421. http://dx.doi.org/10.1002/sce.20303

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Baddeley, A. (2007). *Working memory, thought, and action*. New York, NY: Oxford University Press.

Bambrick-Santoyo, P. (2010). *Driven by data: A practical guide to improve instruction*. San Francisco, CA: Jossey-Bass.

Behrens, J. T., Mislevy, R. J., Bauer, M., Williamson, D. M., & Levy, R. (2004). Introduction to evidence centered design and lessons learned from its application in a global e-learning program. *International Journal of Testing*, *4*(4), 295-301. http://dx.doi.org/10.1207/s15327574ijt0404_1

Birenbaum, M., Kelly, A. E., & Tatsuoka, K. K. (1993). Diagnosing knowledge states in algebra using the rule space model. *Journal of Research in Mathematics Education*, *24*, 442-459. http://dx.doi.org/10.2307/749153

Bjorklund, D. F. (2005). Children's thinking: Cognitive development and individual differences (4th ed.). Belmont, CA: Thomson Learning.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004, September). Working inside the black box: Assessment for learning in the classroom. *Phi Delta Kappan*, *86*(1), 8-21. Retrieved from http://web.ebscohost.com

Black, P., & Wiliam, D. (1998a, March). Assessment and classroom learning. *Assessment in Education: Principles, Policy, & Practice*, *5*(1), 7-68. http://dx.doi.org/10.1080/0969595980050102

Black, P., & Wiliam, D. (1998b, October). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *80*(2), 1-13. Retrieved from http://web.ebscohost.com

Blanc, S., Christman, J. B., Liu, R., Mitchell, C., Travers, E., & Bulkley, K. E. (2010). Learning to learn from data: Benchmarks and instructional communities. *Peabody Journal of Education*, *85*, 205-225. http://dx.doi.org/10.1080/01619561003685379

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Boudett, K. P., City, E. A., & Murnane, R. J. (2010). *Data wise: A step-by-step guide to using assessment results to improve teaching and learning*. Cambridge, MA: Harvard Education Press.

Boudett, K. P., Murnane, R. J., & City, E. (2005). Teaching educators how to use student assessment data to improve instruction. *Phi Delta Kappan*, *86*(9), 700-706. Retrieved from http://scholar.google.com

Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment*, *11*(1), 33-63. Retrieved from http://scholar.google.com

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods and Research*, *21*, 230-258.

Carlson, D., Borman, G. D., & Robinson, M. (2011). A multistate district-level cluster randomized trial of the impact of data-driven reform on reading and mathematics achievement. *Educational Evaluation and Policy Analysis*, *33*(3), 378-398. http://dx.doi.org/10.3102/0162373711412765

Christman, J. B., Neild, R. C., Bulkey, K., Blanc, S., Liu, R., Mitchell, C., & Travers, E. (2009). *Making the most of interim assessment data: Lessons from Philadelphia*. Retrieved from Research for Action website: www.researchforaction.org

Clune, W. H., & White, P. A. (2008). *Policy effectiveness of interim assessments in Providence public schools* (WCER Working Paper No. 2008-10). Retrieved from Wisconsin Center for Education Research website: http://www.wcer.wisc.edu/

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (2nd ed.). Thousand Oaks, CA: Sage.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Mason, OH: Cengage Learning.

Daro, P., Mosher, F. A., & Corcoran, T. (2011). *Learning trajectories in mathematics: A foundation for standards, curriculum, assessment, and instruction* (CPRE Research Report# RR-68). Retrieved from Consortium for Policy Research in Education website: http://cpre.org/

Datnow, A., Park, V., & Wohlstetter, P. (2007). *Achieving with data: How high-performing school systems use data to improve instruction for elementary students*. Retrieved from University of Southern California, Center on Educational Governance website: http://www.uscrossier.org

Dehn, M. J. (2008). Working memory and academic learning: Assessment and intervention. Hoboken, NJ: John Wiley & Sons.

Dehn, M. J. (2010). Long-term memory problems in children and adolescents: Assessment, intervention, and effective instruction. Hoboken, NJ: John Wiley & Sons.

Elementary and Secondary Education Act of 1965, 20 U.S.C. § 6301 *et seq.* (1965).

Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*(1), 179-197. Retrieved from http://psycnet.apa.org

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, *3*(3), 380-396. http://dx.doi.org/10.1037/1082-989X.3.3.380

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika*, *64*(4), 407-433. Retrieved from http://www.psychometrika.org/

Embretson, S. E. (Ed.). (2010). Cognitive design systems: a structural modeling approach applied to developing a spatial ability test. *Measuring psychological constructs: Advances in model-based approaches* (pp. 247-273). http://dx.doi.org/10.1037/12074-011

Embretson, S., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, *38*(4), 343-368. http://dx.doi.org/10.1111/j.1745-3984.2001.tb01131.x

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

Feifer, S. G., & De Fina, P. A. (2005). *The neuropsychology of mathematics: Diagnosis and intervention*. Middletown, MD: School Neuropsych Press.

Geary, D. C., Hoard, M. K., & Bailey, D. H. (2011). Fact retrieval deficits in low achieving children and children with mathematical learning disability. *Journal of Learning Disabilities*, *45*(4), 291-307. http://dx.doi.org/10.1177/0022219410392046

Gierl, M. J. (2007). Making diagnostic inferences about cognitive attributes using the rule-space model and attribute hierarchy method. *Journal of Educational Measurement*, *44*(4), 325-340. http://dx.doi.org/10.1111/j.1745-3984.2007.00042.x

Gierl, M. J., Alves, C., & Majeau, R. T. (2010). Using the attribute hierarchy method to make diagnostic inferences about examinees' knowledge and skills in mathematics: An operational implementation of cognitive diagnostic assessment. *International Journal of Testing*, *10*, 318-341. http://dx.doi.org/10.1080/15305058.2010.509554

Gierl, M. J., Leighton, J. P., & Hunka, S. M. (2000). Exploring the logic of Tatsuoka's rule-space model for test development and analysis. *Educational Measurement: Issues and Practice*, *19*(3), 34-44. http://dx.doi.org/10.1111/j.1745-3992.2000.tb00036.x

Gierl, M. J., Leighton, J. P., & Hunka, S. M. (2007). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills. In J. P. Leighton,

& M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 242-274). New York, NY: Cambridge University Press.

Goertz, M. E., Olah, L. N., & Riggan, M. (2009). *From testing to teaching: The use of interim assessments in classroom instruction* (CPRE Research Report No. RR-65). Retrieved from Consortium for Policy Research in Education website: http://www.cpre.org

Goren, P. (2010). Interim assessments as a strategy for improvement: Easier said than done. . *Peabody Journal of Education*, *85*, 125-129. http://dx.doi.org/10.1080/01619561003673938

Gorin, J. S. (2007). Test construction and diagnostic testing. In J. P. Leighton, & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 173-201). New York, NY: Cambridge University Press.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. Mahwah, NJ: Lawrence Erlbaum Associates.

Haladyna, T. M., & Downing, S. M. (1989a). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, *2*(1), 37-50.

Haladyna, T. M., & Downing, S. M. (1989b). Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, *2*(1), 51-78.

Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement*, *53*, 999-1010. http://dx.doi.org/10.1177/0013164493053004013

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. New York, NY: Routledge.

Hamilton, L. S., Nussbaum, E. M., & Snow, R. E. (1997). Interview procedures for validating science assessments. *Applied Measurement in Education*, *10*(2), 181-201. Retrieved from http://web.ebscohost.com

Henderson, S., Petrosino, A., Guckenburg, S., & Hamilton, S. (2007). *Measuring how benchmark assessments affect student achievement* (REL2007- No. 039). Retrieved from Institute of Education Sciences: http://ies.ed.gov/

Hermann-Abell, C. F., & DeBoer, G. E. (2011). Using distractor-driven stndards-based multiple-choice assessments and Rasch modeling to investigate hierarchies of chemistry misconceptions and detect structural problems with individual items. *Chemistry Education Research and Practice*, *12*, 184-192. http://dx.doi.org/10.1039/C1RP90023D

Hu, W. (2011, December 11). State takeovers of other districts have had mixed results. *New York Times*. Retrieved from http://www.nytimes.com/

Huff, K., & Goodman, D. P. (2007). The demand for cognitive diagnostic assessment. In J. P. Leighton, & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 19-60). New York, NY: Cambridge University Press.

IBM SPSS Statistics for Windows (Version 22.0) [Computer software]. (2013). Armonk, NY: IBM Corp.

Kamphaus, R. W. (2005). *Clinical assessment of child and adolescent intelligence* (2nd ed.). New York, NY: Springer.

Kane, M. (2006). Content-related validity evidence in test development. In S. M. Downing, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 131-153). Mahwah, NJ: Lawrence Erlbaum Associates.

Kane, M. (2009). Validating the interpretations and uses of test scores. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 39-64). Charlotte, NC: Information Age .

Kruschke, J. K. (2005). Category learning. In K. Lamberts, & R. L. Goldstone (Eds.), *Handbook of Cognition* (pp. 183-201). Thousand Oaks, CA: Sage.

Leighton, J. P., & Gierl, M. J. (2007a). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice*, *26*(2), 3-16. http://dx.doi.org/10.1111/j.1745-3992.2007.00090.x

Leighton, J. P., & Gierl, M. J. (Eds.). (2007b). *Cognitive diagnostic assessment for education: Theory and applications*. New York, NY: Cambridge University Press.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement*, *41*(3), 205-237. http://dx.doi.org/10.1111/j.1745-3984.2004.tb01163.x

Levine, M. V., & Drasgow, F. (1983). The relation between incorrect option choice and estimated ability. *Educational and Psychological Measurement*, *43*, 675-687. http://dx.doi.org/10.1177/001316448304300301

Li, Y., Marion, S., Perie, M., & Gong, B. (2010). An approach for evaluating the technical quality of interim assessments. *Peabody Journal of Education*, *85*, 163-185. http://dx.doi.org/10.1080/01619561003685304

Livingston, S. A. (2006). Item analysis. In S. M. Downing, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 421-441). Mahwah, NJ: Lawrence Erlbaum Associates.

Lord, F. M. (1977, Spring). Optimal number of choices per item: A comparison of four approaches. *Journal of Educational Measurement*, *14*(1), 33-38. Retrieved from http://www.jstor.org

Mandinach, E. B., & Honey, M. (2008). *Data-driven school improvement: Linking data and learning*. New York, NY: Teachers College Press.

Marsh, J. A., Pane, J. F., & Hamilton, L. S. (2006). *Making sense of data-driven decision making in education* [Occasional paper]. Retrieved from RAND website: http://www.rand.org

Marshall, K. (2008). Interim assessments: A user's guide. *Phi Delta Kappan*, *90*(1), 64-68. Retrieved from http://intl.kappanmagazine.org

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174. Retrieved from http://www.psychometrika.org/

Matlin, M. W. (2002). *Cognition* (5th ed.). Geneseo, NY: Thomson Learning.

Mazzocco, M. M., & Devlin, K. T. (2008). Parts and 'holes': Gaps in rational number sense among children with vs. without mathematical learning disabilities. *Developmental Science*, *11*(5), 681-691. http://dx.doi.org/10.1111/j.1467-7687.2008.00717.x

Mazzocco, M. M., Myers, G. F., Lewis, K. E., Hanich, L. B., & Murphy, M. M. (2013). Limited knowledge of fraction representations differentiates middle school students with mathematics learning disability (dyscalculia) versus low mathematics achievement. *Jounal of Experimental Child Psychology*, *115*, 317-387. http://dx.doi.org/10.1016/j.jecp.2013.01.005

Meyer, J. P. (2002). jMetrik (Version 3.0.1) [Computer program]. Retrieved from http://itemanalysis.com/

Meyer, J. P. (in press). *Applied measurement with jMetrik*. New York, NY: Routledge.

Miles, M. B., & Huberman, A. M. (1994). Matrix displays: Some rules of thumb. In *Qualitative data analysis* (2nd ed., pp. 239-244). Thousand Oaks, CA: Sage.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centered design* (ETS Report No. RR-03-16). Retrieved from Educational Testing Service website: http://www.ets.org

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, *25*(4), 6-20. Retrieved from http://web.ebscohost.com

Naglieri, J. A., & Goldstein, S. (2013). Comprehensive Executive Function Inventory [Technical Manual]. Published instrument. Retrieved from www.mhs.com

National Research Council. (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: National Academy Press.

Nichols, P. D. (1994). A framework for developing cognitively diagnostic assessments. *Review of Educational Research*, *64*, 575-603. Retrieved from http://www.jstor.org/

No Child Left Behind Act of 2001, 20 U.S.C. § 6301 *et seq.* (2002).

Oláh, L. N., Lawrence, N. R., & Riggan, M. (2010). Learning to learn from benchmark assessment data: How teachers analyze results. *Peabody Journal of Education*, *85*, 226-245. http://dx.doi.org/10.1080/01619561003688688

Perie, M., Marion, S., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practices*, *28*(3), 5-13.

Perie, M., Marion, S., Gong, B., & Wurtzel, J. (2007). *The role of interim assessments in a comprehensive assessment system: A policy brief* [Policy brief]. Retrieved from National Center for the Improvement of Educational Assessment website: www.nciea.org

Rittle-Johnson, B., & Siegler, R. S. (1998). The relation between conceptual and procedural knowledge in learning mathematics: A review. In C. Donlan (Ed.), *The development of mathematical skills* (pp. 75-110). East Sussex, United Kingdom: Psychology Press.

Rundquist, J. (2013, September 5). Christie to Newark: We run the school district. *The Star-Ledger*. Retrieved from http://www.nj.com/

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guildford Press.

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement*, *6*, 219-262. http://dx.doi.org/10.1080/15366360802490866

Sadler, P. M. (1998). Psychometric models of student conceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments. *Journal of Research in Science Teaching*, *35*(3), 265-296. http://dx.doi.org/10.1002/(SICI)1098-2736(199803)35

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307-353). Westport, CT: Praeger.

School boards file lawsuit claiming Virginia school takeover law is unconstitutional. (2013, September 13). *Washington Post*. Retrieved from http://www.washingtonpost.com/

Sergeant, J. (1996). A theory of attention: An information processing perspective. In G. R. Lyon, & N. A. Krasnegor (Eds.), *Attention, memory, and executive function* (pp. 57-69). Baltimore, MD: Paul H. Brookes.

Shepard, L. A. (2010). What the marketplace has brought us: Item-by-item teaching with little instructional insight. *Peabody Journal of Education*, *85*, 246-257. http://dx.doi.org/10.1080/01619561003708445

Steinberg, L. S., Mislevy, R. J., Almond, R. G., Baird, A. B., Cahallan, C., Dibello, L. V.,...Kindfield, A. C. (2003). *Introduction to the biomass project: An illustration of evidence-centered assessment design and delivery capability* (CSE-R 609). Retrieved from CRESST website: http://www.cse.ucla.edu/

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*(4), 345-354. Retrieved from http://www.jstor.org/

Tatsuoka, K. K. (1986). Diagnosing cognitive errors: Statistical pattern classification based on item response theory. *Behaviormetrika*, *13*, 73-86. http://dx.doi.org/10.2333/bhmk.13.19_73

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453-488). Retrieved from www.books.google.com

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern and classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327-359). Retrieved from http://books.google.com

Tatsuoka, K. K. (2009). Cognitive assessment: An introduction to the rule space method. New York, NY: Routledge.

United States Department of Education. (2010). *ESEA blueprint for reform* [Policy brief]. Retrieved from United States Department of Education website: http://www2.ed.gov/

Virginia Department of Education. (2009). *Mathematics Standards of Learning curriculum framework 2009: Grade 7* [Educational standards]. Retrieved from VDOE website: http://www.doe.virginia.gov/

Virginia Department of Education. (2009). *Virginia Standards of Learning assessments test blueprint: Grade 7 mathematics 2009 mathematics Standards of Learning* [Educational standards]. Retrieved from VDOE website: http://www.doe.virginia.gov/

Virginia Department of Education website. (n.d.). http://www.doe.virginia.gov/

Wiliam, D. (2011). *Embedded formative assessment*. Bloomington, IN: Solution Tree Press.

Williamson, D. M., Bauer, M., Steinberg, L. S., Mislevy, R. J., Behrens, J. T., & DeMark, S. F. (2004). Design rationale for a complex performance assessment. *International Journal of Testing*, *4*(4), 303-332. http://dx.doi.org/10.1207/s15327574ijt0404_2

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wilson, M. (2008). Cognitive diagnosis using item response models. *Journal of Psychology*, *216*(2), 74-88. http://dx.doi.org/10.1027/0044-3409.216.2.74

Young, V. M. (2006). Teachers's use of data: Loose coupling, agenda setting, and team norms. *American Journal of Education*, *112*, 521-548. Retrieved from http://www.jstor.org

Appendix A: Mathematics 6A: Student item responses for the second nine weeks interim assessment

Item#	SOL	Answer	A	B	C	D	Missing
1	7.1g	C	26 - 10.5%	41 - 16.5%	**165 - 66.5%**	16 - 6.5%	0 - 0.0%
2	7.3c	D	3 - 1.3%	90 - 36.3%	8 - 3.2%	**147 - 59.3%**	0 - 0.0%
3	7.3c	A	**201 - 81.1%**	19 - 7.7%	19 - 7.7%	9 - 3.6%	0 - 0.0%
4	7.1f	B	95 - 38.4%	**133 - 53.6%**	4 - 1.6%	15 - 6.0%	1 - 0.4%
5	7.4b	A	**168 - 67.8%**	3 - 1.2%	50 - 20.2%	26 - 10.5%	1 - 0.4%
6	7.4c	D	37 - 15.0%	50 - 20.2%	43 - 17.3%	**118 - 47.6%**	0 - 0.0%
7	7.3a	C	24 - 9.7%	29 - 11.7%	**133 - 53.6%**	62 - 25.0%	0 - 0.0%
8	7.4a	A	**80 - 32.3%**	56 - 22.6%	39 - 15.7%	73 - 29.4%	0 - 0.0%
9	7.7a	C	11 - 4.5%	20 - 8.1%	**144 - 58.1%**	73 - 29.4%	0 - 0.0%
10	7.4d	D	24 - 9.7%	32 - 12.9%	22 - 8.9%	**170 - 68.5%**	0 - 0.0%
11	7.7b	B	9 - 3.7%	**13 - 5.2%**	208 - 83.9%	18 - 7.3%	0 - 0.0%
12	7.8c	B	25 - 10.1%	**126 - 50.8%**	82 - 33.1%	15 - 6.0%	0 - 0.0%
13	7.6b	A	**57 - 23.0%**	35 - 14.1%	31 - 12.5%	125 - 50.4%	0 - 0.0%
14	7.8f	D	48 - 19.4%	20 - 8.1%	14 - 5.6%	**165 - 66.5%**	1 - 0.4%
15	7.8e	A	**105 - 42.4%**	17 - 6.9%	31 - 12.5%	94 - 37.9%	1 - 0.4%
16	7.5c	B	25 - 10.1%	**53 - 21.4%**	133 - 53.6%	37 - 14.9%	0 - 0.0%
17	7.4f	C	2 - 0.9%	51 - 20.6%	**139 - 56.0%**	56 - 22.6%	0 - 0.0%
18	7.9b	D	8 - 3.3%	26 - 10.5%	69 - 27.8%	**145 - 58.5%**	0 - 0.0%
19	7.9a	A	**216 - 87.1%**	4 - 1.6%	27 - 10.9%	1 - 0.4%	0 - 0.0%
20	7.5i	C	65 - 26.3%	15 - 6.0%	**153 - 61.7%**	14 - 5.6%	1 - 0.4%
21	7.5j	B	46 - 18.6%	**156 - 62.9%**	33 - 13.3%	13 - 5.2%	0 - 0.0%
22	7.5d	B	19 - 7.7%	**163 - 65.7%**	27 - 10.9%	39 - 15.7%	0 - 0.0%
23	7.5f	C	5 - 2.1%	62 - 25.0%	**178 - 71.8%**	3 - 1.2%	0 - 0.0%
24	7.6d	D	22 - 8.9%	27 - 10.9%	13 - 5.2%	**186 - 75.0%**	0 - 0.0%
25	7.6a	D	105 - 42.4%	16 - 6.5%	36 - 14.5%	**88 - 35.5%**	3 - 1.2%

**Quantities represent the number and percent of students who selected a specific response option. Correct responses are in bold.*

Appendix B: Mathematics 7R: Student item responses for the second nine weeks interim assessment

Item #	SOL	Correct Answer	A	B	C	D	Missing
1	7.1g	C	40 - 14.4%	75 - 27.0%	**130 - 46.8%**	30 - 10.8%	3 - 1.1%
2	7.3c	D	11 - 4.0%	106 - 38.1%	10 - 3.6%	**149 - 53.6%**	2 - 0.7%
3	7.3c	A	**202 - 72.7%**	47 - 16.9%	21 - 7.6%	6 - 2.2%	2 - 0.7%
4	7.1f	B	135 - 48.6%	**103 - 37.1%**	22 - 7.9%	15 - 5.4%	3 - 1.1%
5	7.4b	A	**127 - 45.7%**	25 - 9.0%	82 - 29.5%	41 - 14.7%	3 - 1.1%
6	7.4c	D	66 - 23.8%	76 - 27.3%	59 - 21.2%	**71 - 25.5%**	6 - 2.2%
7	7.3a	C	30 - 10.8%	37 - 13.3%	**140 - 50.4%**	69 - 24.8%	2 - 0.7%
8	7.4a	A	**64 - 23.1%**	91 - 32.7%	83 - 29.9%	37 - 13.3%	3 - 1.1%
9	7.7a	C	27 - 9.8%	51 - 18.3%	**101 - 36.3%**	98 - 35.3%	1 - 0.4%
10	7.4d	D	54 - 19.5%	83 - 29.9%	37 - 13.3%	**104 - 37.4%**	0 - 0.0%
11	7.7b	B	29 - 10.5%	**35 - 12.6%**	174 - 62.6%	40 - 14.4%	0 - 0.0%
12	7.8c	B	34 - 12.3%	**112 - 40.3%**	97 - 34.9%	35 - 12.6%	0 - 0.0%
13	7.6b	A	**76 - 27.4%**	49 - 17.6%	28 - 10.1%	123 - 44.2%	2 - 0.7%
14	7.8f	D	59 - 21.3%	30 - 10.8%	40 - 14.4%	**149 - 53.6%**	0 - 0.0%
15	7.8e	A	**89 - 32.1%**	67 - 24.1%	43 - 15.5%	79 - 28.4%	0 - 0.0%
16	7.5c	B	66 - 23.8%	**48 - 17.3%**	102 - 36.7%	60 - 21.6%	2 - 0.7%
17	7.4f	C	12 - 4.4%	86 - 30.9%	**98 - 35.3%**	81 - 29.1%	1 - 0.4%
18	7.9b	D	38 - 13.7%	41 - 14.7%	102 - 36.7%	**96 - 34.5%**	1 - 0.4%
19	7.9a	A	**233 - 83.9%**	11 - 4.0%	25 - 9.0%	9 - 3.2%	0 - 0.0%
20	7.5i	C	108 - 38.9%	39 - 14.0%	**102 - 36.7%**	27 - 9.7%	2 - 0.7%
21	7.5j	B	94 - 33.9%	**88 - 31.7%**	67 - 24.1%	29 - 10.4%	0 - 0.0%
22	7.5d	B	33 - 11.9%	**137 - 49.3%**	45 - 16.2%	62 - 22.3%	1 - 0.4%
23	7.5f	C	17 - 6.2%	79 - 28.4%	**173 - 62.2%**	9 - 3.2%	0 - 0.0%
24	7.6d	D	33 - 11.9%	45 - 16.2%	39 - 14.0%	**160 - 57.6%**	1 - 0.4%
25	7.6a	D	126 - 45.4%	32 - 11.5%	66 - 23.7%	**53 - 19.1%**	1 - 0.4%

**Quantities represent the number and percent of students who selected a specific response option. Correct responses are in bold.*

Appendix C: Item Scoring and Partial Credit Recoding

Item #	Correct Answer	Item Scoring w/Error Codes	Items w/Partial Credit ReCoding*
1	C	A=0, B=1,C=3,D=2	
2	B	A=0, B=3,C=2,D=1	
3	D	A=2, B=0,C=1,D=3	
4	B	A=2, B=3,C=1,D=1	A=1, B=2,C=0,D=0
5	A	A=3, B=0,C=0,D=0	A=1, B=0,C=0,D=0
6	B	A=1, B=3,C=1,D=0	A=1, B=2,C=1,D=0
7	A	A=3, B=0,C=0,D=2	A=2, B=0,C=0,D=1
8	C	A=0, B=1,C=3,D=1	A=0, B=1,C=2,D=1
9	A	A=3, B=2,C=1,D=1	A=2, B=1,C=0,D=0
10	B	A=0, B=3,C=0,D=2	A=0, B=2,C=0,D=1
11	D	A=0, B=0,C=2,D=3	A=0, B=0,C=1,D=2
12	A	A=3, B=0,C=0,D=1	A=2, B=0,C=0,D=1
13	C	A=0, B=2,C=3,D=1	
14	D	A=1, B=1,C=2,D=3	A=0, B=0,C=1,D=2
15	C	A=1, B=0,C=3,D=0	A=1, B=0,C=2,D=0
16	C	A=2, B=0,C=3,D=1	
17	A	A=3, B=2,C=1,D=1	A=2, B=1,C=0,D=0
18	B	A=0, B=3,C=2,D=1	
19	C	A=0, B=1,C=3,D=0	A=0, B=1,C=2,D=0
20	D	A=1, B=1,C=2,D=3	A=0, B=0,C=1,D=2
21	C	A=1, B=0,C=3,D=2	
22	A	A=3, B=2,C=1,D=0	
23	D	A=0, B=0,C=0,D=3	A=0, B=0,C=0,D=1
24	B	A=1, B=3,C=1,D=0	A=1, B=2,C=1,D=0
25	C	A=0, B=1,C=3,D=0	A=0, B=1,C=2,D=0

**Error Categories are collapsed to allow for Partial Credit Modeling. Items with 3 assigned error categories did not need to be recoded.*

Appendix D: Classical Test Theory: Item Analysis Statistics

Item #	Item Difficulty	Standard Deviation	Item Discrimination**
1	1.9730	1.1516	0.3350
2	2.5541	0.8641	0.2881
3	2.3649	0.5598	0.2691
4	1.3745	0.9070	0.2240
5	0.6931*	0.4617	0.2937
6	1.1313	0.7807	0.0914
7	0.6988	0.9443	0.2702
8	1.0212	0.7551	0.1751
9	1.2645	0.8672	0.0934
10	0.8842	0.8938	0.2732
11	0.9440	0.6364	0.2517
12	1.0676	0.8847	0.1612
13	1.5328	1.0564	0.1662
14	0.8494	0.9101	0.1452
15	1.0039	0.9246	0.2647
16	1.0734	1.1941	-0.0174
17	0.9961	0.6493	0.2916
18	2.2548	1.0613	0.3625
19	1.4131	0.7004	0.2952
20	0.9537	0.8538	0.2678
21	1.8822	1.1881	0.3241
22	2.3938	0.9326	0.4079
23	0.3166*	0.4656	0.0521
24	1.4595	0.6261	0.2618
25	0.7471	0.9533	0.2462

**Dichotomously scored items. All distractors represent conceptual errors with assigned score= 0*

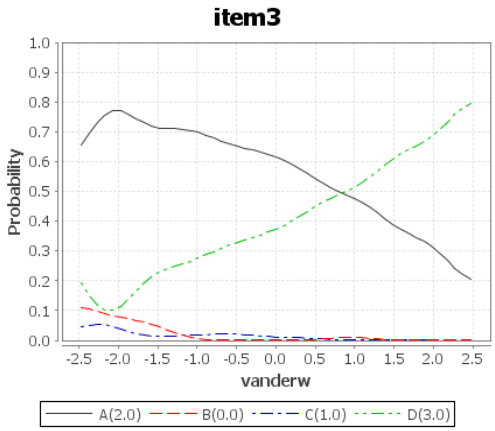
***Polyserial Correlation*

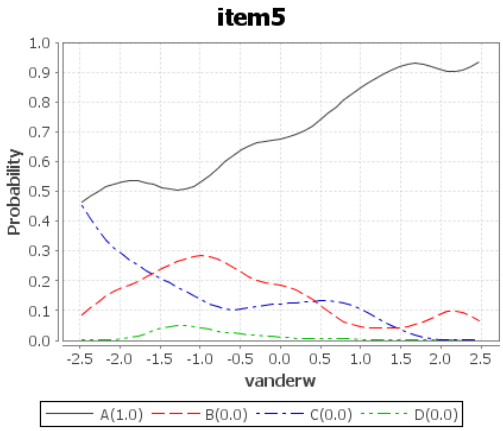
Appendix E: Distractor Frequency Analysis (%) for Partial Credit Items

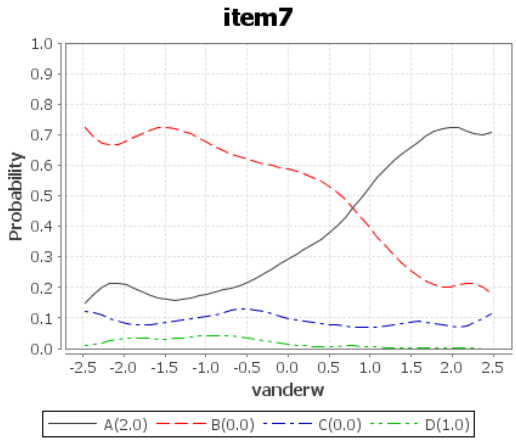
Item #	A	B	C	D	Missing
1	15.1	21.2	**49.4***	13.9	0.4
2**	4.8	**75.5**	9.3	10.4	0
3**	58.3	1.0	1.2	**39.6**	0
4**	3.9	**66.8**	11.8	17.4	0.2
5**	**69.3**	16.2	12.5	1.5	0.4
6	15.6	**37.8**	21.8	23.9	0.8
7**	**34.0**	54.4	9.5	1.9	0.2
8	27.2	25.9	**29.5**	17.2	0.2
9**	**54.2**	18.0	23.6	3.9	0.4
10	28.6	**34.7**	17.2	18.9	0.6
11	11.6	11.4	59.3	**17.6**	0.2
12	**42.7**	23.9	11.6	21.4	0.4
13	14.3	10.8	**28.6**	45.9	0.4
14	25.7	23.6	15.1	**34.9**	0.8
15	14.7	14.5	**42.9**	27.8	0.2
16	17.8	48.1	**19.1**	14.5	0.6
17	**20.8**	57.9	13.9	6.6	0.8
18	10.4	**61.2**	14.1	13.7	0.6
19	6.4	34.0	**53.7**	5.6	0.4
20	17.6	20.7	27.0	**34.2**	0.6
21	10.4	21.4	**43.1**	24.3	0.8
22	**66.6**	10.4	18.7	4.2	0
23	18.1	28.0	27.2	**31.7**	0
24	25.7	**53.1**	14.1	6.9	0.2
25**	42.3	2.9	**35.9**	18.5	0.4

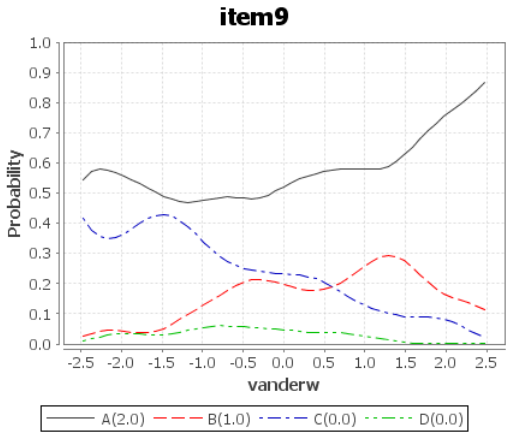
**Correct responses are in bold.*
***Distractors with < 5% response rate*

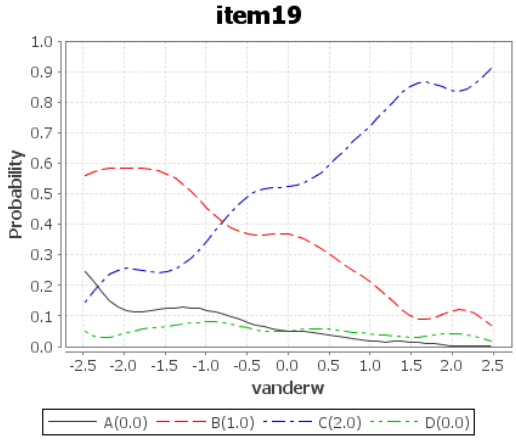
Appendix F: Items with Flat Nonparametric Curves

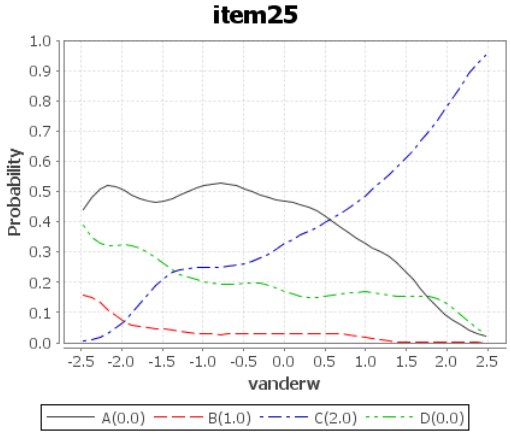




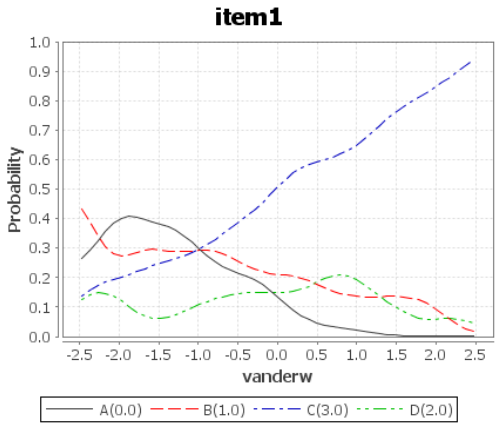


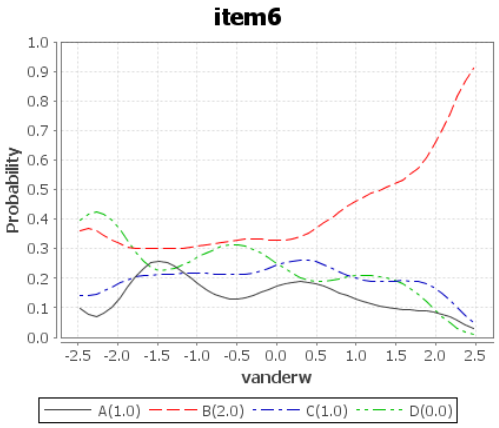


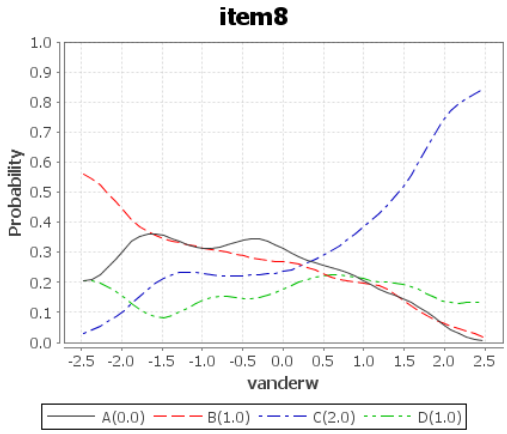


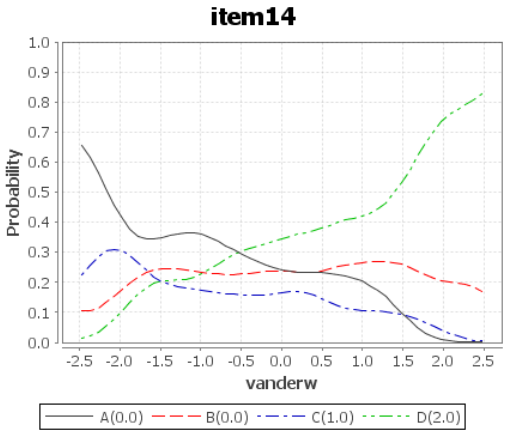


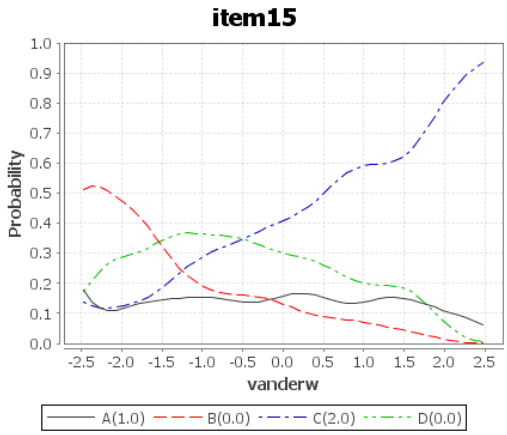
Appendix G: Items with Non-discriminating Nonparametric Curves

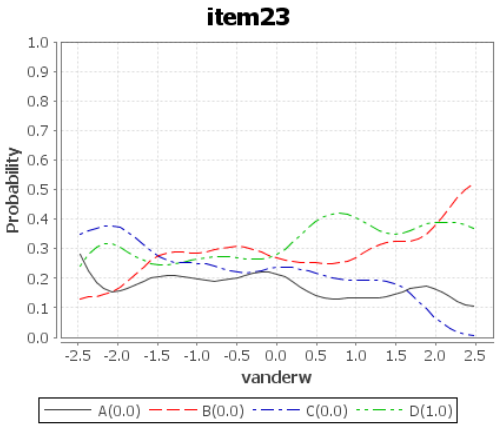












Appendix H: IRT Item Parameter and Fit Statistics

Item #	Difficulty	Std Error	WMS (infit)	UMS (outfit)
1	-0.18	0.04	0.94	0.90
2	-0.76	0.05	0.92	0.88
3	-1.04	0.08	0.95	0.94
4	-0.19	0.05	1.00	1.01
5	-0.56	0.10	0.95	0.93
6	0.02	0.06	1.06	1.06
7	0.60	0.05	0.99	1.01
8	0.22	0.06	1.01	1.00
9	-0.11	0.05	1.07	1.06
10	0.41	0.05	0.98	0.97
11	0.40	0.07	0.97	0.97
12	0.16	0.05	1.04	1.05
13	0.03	0.05	1.07	1.07
14	0.44	0.05	1.06	1.06
15	0.25	0.05	0.99	0.98
16	0.61	0.04	1.26	1.39
17	0.26	0.07	0.94	0.94
18	-0.41	0.04	0.90	0.85
19	-0.52	0.07	0.94	0.93
20	0.32	0.05	0.98	0.97
21	-0.02	0.04	0.96	0.91
22	-0.74	0.05	0.87	0.81
23	1.03	0.10	1.03	1.03
24	-0.78	0.07	0.95	0.95
25	0.55	0.05	1.01	1.00

Appendix I: IRT Item Category Thresholds for Items with 3 Error Categories

Item #	Category	Thresholds	SE	WMS	UMS
1*	0				
	1	-0.08	0.13	1.01	1.01
	2	0.81	0.10	0.83	0.73
	3	-0.74	0.09	0.98	0.98
2*	0				
	1	0.02	0.20	0.99	1.00
	2	1.03	0.12	0.89	0.64
	3	-1.05	0.11	0.97	0.99
3*	0				
	1	0.92	0.43	0.91	0.89
	2	-2.69	0.30	0.94	0.93
	3	1.77	0.09	0.99	1.00
13*	0				
	1	-1.08	0.13	1.15	1.12
	2	1.67	0.10	1.24	1.22
	3	-0.59	0.10	1.05	1.05
16*	0				
	1	0.75	0.09	1.12	1.98
	2	-0.52	0.10	1.19	1.24
	3	-0.23	0.12	1.14	1.20
18*	0				
	1	0.19	0.15	0.96	0.97
	2	0.55	0.11	0.74	0.55
	3	-0.75	0.09	0.91	0.94
21*	0				
	1	0.82	0.11	0.64	0.52
	2	-0.62	0.10	0.86	0.72
	3	-0.20	0.09	0.96	0.97
22*	0				
	1	-0.69	0.21	0.86	0.77
	2	1.50	0.11	0.82	0.64
	3	-0.80	0.10	0.90	0.94

**Items with Reversals.* *SE = Standard Error*

Appendix I (cont'd): IRT Item Category Thresholds for Items with Conceptual and Procedural Errors

Item #	Category	Thresholds	SE	WMS	UMS
6	0				
	1	-0.30	0.11	0.91	0.83
	2	0.30	0.09	1.06	1.06
8	0				
	1	-0.50	0.10	0.94	0.90
	2	0.50	0.10	0.97	0.97
12*	0				
	1	0.52	0.10	1.03	1.00
	2	-0.52	0.09	1.03	1.03
15*	0				
	1	0.99	0.09	1.09	1.05
	2	-0.99	0.09	0.97	0.97
19	0				
	1	-0.37	0.14	0.81	0.74
	2	0.37	0.09	0.93	0.95
24	0				
	1	-0.80	0.17	0.88	0.82
	2	0.80	0.09	0.97	0.98
25*	0				
	1	2.70	0.10	1.49	1.73
	2	-2.70	0.10	0.98	0.98

**Items with Reversals.* *SE = Standard Error*

IRT Item Category Thresholds for Items with Conceptual and Attention Errors

Item #	Category	Thresholds	SE	WMS	UMS
7*	0				
	1	3.10	0.10	1.55	1.48
	2	-3.10	0.10	0.96	0.97
10*	0				
	1	0.67	0.09	0.97	0.92
	2	-0.67	0.10	0.97	0.97
11	0				
	1	-1.17	0.11	1.01	1.01
	2	1.17	0.12	0.99	0.99

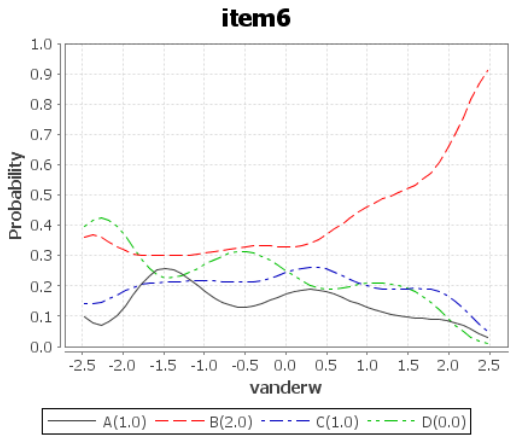
**Items with Reversals.* *SE = Standard Error*

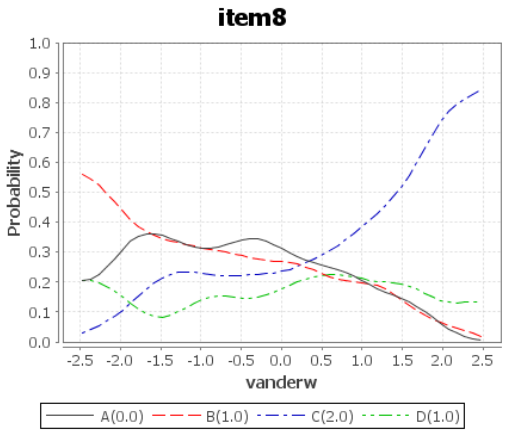
Appendix I (cont'd): IRT Item Category Thresholds for Items with Attention and Procedural Errors

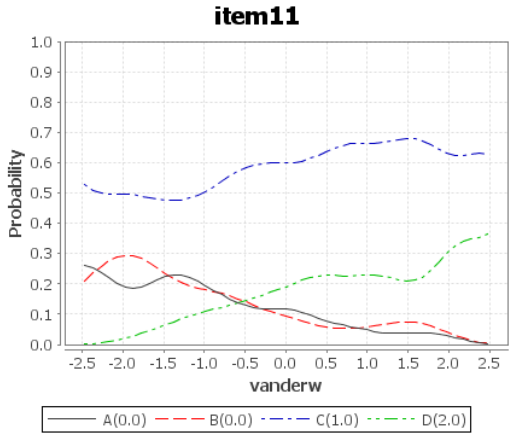
Item #	Category	Thresholds	SE	WMS	UMS
4*	0				
	1	2.37	0.10	0.48	0.60
	2	-2.37	0.10	1.00	1.02
9*	0				
	1	0.69	0.10	1.36	1.37
	2	-0.69	0.09	1.13	1.11
14*	0				
	1	0.94	0.09	1.16	1.11
	2	-0.94	0.10	1.00	1.00
17	0				
	1	-1.11	0.11	0.89	0.86
	2	1.11	0.11	0.92	0.92
20*	0				
	1	0.21	0.09	1.12	1.16
	2	-0.21	0.10	0.94	0.93

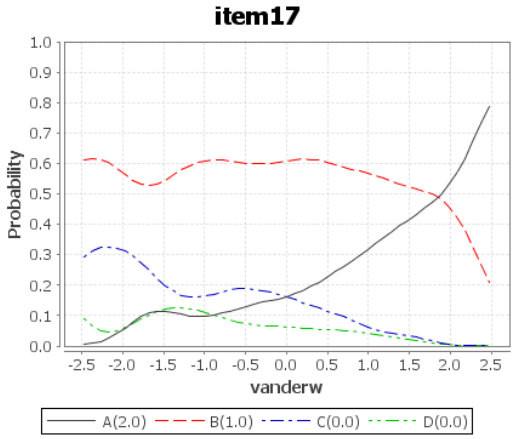
**Items with Reversals* *SE = Standard Error*

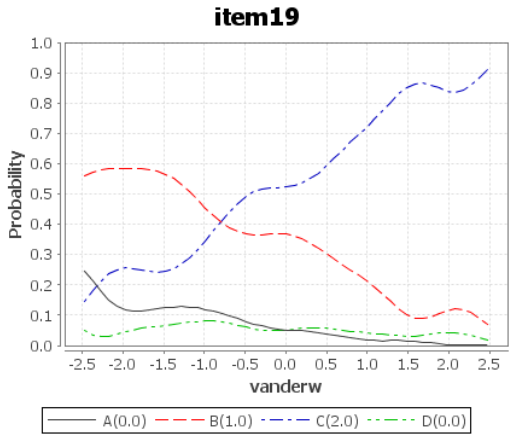
Appendix J: Nonparametric Curves of Items Without Reversals

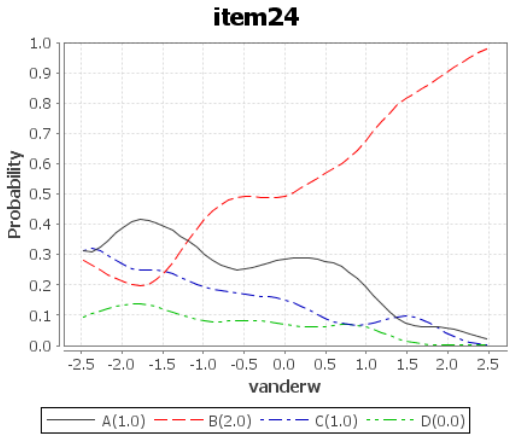












Appendix K: Item Map