
Ally Thomas
The Graduate Center, City University of New York
EVALUATING THE VALIDITY OF TECHNOLOGY-ENHANCED EDUCATIONAL ASSESSMENT ITEMS AND TASKS:
AN EMPIRICAL APPROACH TO STUDYING ITEM FEATURES AND SCORING RUBRICS

by

ALLY THOMAS

A dissertation submitted to the Graduate Faculty in Educational Psychology in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

2016

by

Ally Thomas

This manuscript has been read and accepted for the Graduate Faculty in Educational Psychology in satisfaction of the dissertation requirement for the Doctor of Philosophy.

Date

Jay Verkuilen, PhD

Chair of Examining Committee

Date

Bruce Homer, PhD

Executive Officer

Supervisory Committee:

Jay Verkuilen, PhD

Howard Everson, PhD

Bruce Homer, PhD

THE CITY UNIVERSITY OF NEW YORK
ABSTRACT

This manuscript has been read and accepted for the Graduate Faculty in Educational Psychology in satisfaction of the dissertation requirement for the Doctor of Philosophy.

by

Ally Thomas

Advisor: Jay Verkuilen, PhD

With the advent of the newly developed Common Core State Standards and the Next Generation Science Standards, innovative assessments, including technology-enhanced items and tasks, will be needed to meet the challenges of developing valid and reliable assessments in a world of computer-based testing. In a recent critique of the next generation assessments in math (i.e., Smarter Balanced), Rasmussen (2015) observed that many aspects of the technology “enhancements” can be expected to do more harm than good as the computer interfaces may introduce construct irrelevant variance. This paper focused on issues surrounding the design of TEIs and how cognitive load theory (Miller, 1956) is a promising framework that can be applied to computer-based item design to mitigate the effects of computer interface usability. Two studies were conducted. In the first study I used multi-level modeling to assess the effect of item characteristics on examinees’ relative performance. I hypothesized that item level characteristics, namely response format, would significantly contribute to the amount of variance explained by item characteristics over and above student characteristics. In study two, I used two exemplar items to show how data concerning examinees’ actions—produced through latent class analyses—can be used as evidence in validity investigations. Results from study 1 suggested that item type does not explain the variation in student scores over and above examinee characteristics. Results from study two suggested that LCA is a useful tool for diagnosing potential issues in the design of items and the design of their scoring rubrics.
Evidence provided from both studies illuminates the immediate need to further research computer-based items that are beginning to be used widely in high stakes, large-scale assessments. In an effort to move away from traditional multiple choice items and toward more authentic measurement by incorporating technology based item features, we may be affecting how examinees respond to the item due to inadvertent increases in cognitive load. Future research involving experimental manipulation is necessary for understanding how item characteristics impact how examinees responses.
# Table of Contents

Chapter 1: Introduction ........................................................................................................ 1

Chapter 2: Literature Review ................................................................................................. 11
  Evolution of Technology in Educational Assessments .................................................... 11
  Contemporary Views on Assessment Design ................................................................. 13
  Current Understanding of Item Design ....................................................................... 17
  The Role of Cognitive Psychology in Assessment Design .......................................... 20
  Hypotheses ....................................................................................................................... 27

Chapter 4. Study Design ...................................................................................................... 28
  Data .................................................................................................................................. 28
  Procedures for Conducting Field Study ....................................................................... 32
  Participants ....................................................................................................................... 34

Chapter 5. Study 1 ................................................................................................................. 36
  Method ............................................................................................................................. 36
  Results ............................................................................................................................. 38
  Summary of Findings ...................................................................................................... 40

Chapter 6. Study 2 ................................................................................................................. 40
  Method ............................................................................................................................. 40
  Results ............................................................................................................................. 43
  Summary of Findings ...................................................................................................... 50

Chapter 7. Discussion ........................................................................................................... 51

Chapter 8. Conclusion ......................................................................................................... 53
  Limitations ....................................................................................................................... 54
  Future Research ............................................................................................................. 54

Appendix A. Task Models for Design of Mathematics Prototype Items ......................... 56

Appendix B. Mapping of Prototype Tasks onto CCSS and Mathematical Practices ........ 58

Appendix C. Study 1 Statistical Model ............................................................................... 60

Appendix D. Study 1 Syntax for Multi-Level Model ............................................................. 61

Appendix E. Study 2 Syntax for Latent Class Analyses ....................................................... 62

References .......................................................................................................................... 63
List of Tables

Table 1. Cognitive Load Effects (Sweller, 2010). ................................................................. 23
Table 2. Distribution of Examinees Tested by Grade Level. .............................................. 34
Table 3. Demographic Characteristics. .................................................................................. 35
Table 4. Examinee Demographic Characteristics for Exemplar Items. ........................... 44
Table 5. Model Fit Indices for Classes 2 through 6. ......................................................... 46
Table 6. Probability of Selecting each Response Option by Latent Class. ....................... 46
Table 7. Frequency of Response Patterns by Class. ............................................................ 47
Table 8. Comparison of Latent Class Membership and Points Earned in Original and Adapted Rubrics. ............................................................................................................. 49
Table 9. Model Fit Indices for Classes 2 through 3. ............................................................. 50
List of Figures

Figure 1. TEI example from WestEd’s SimScientists (Quellmalz, DeBarger, Haertel, Schank, Buckley, et al., 2008) ................................................................. 3

Figure 2. TEI example of ETS’s Cognitively Based Assessment of, for, and as Learning (CBAL, Bennett & Gitomer, 2009) ................................................................. 4

Figure 3. Energy Transfer NAEP science activity part 1 (NCES, 2012) .................................................. 6

Figure 4. Energy Transfer NAEP science activity part 2 (NCES, 2012) .................................................. 7

Figure 5. Energy Transfer NAEP science activity part 3 (NCES, 2012) .................................................. 8

Figure 6. Energy Transfer NAEP science activity part 4 (NCES, 2012) .................................................. 9

Figure 7. Structure for assessment arguments (Mislevy, 2008) ................................................................. 15

Figure 8. Grade 3 Mathematics Prototype Item (http://ccsstoolbox.agilemind.com) ......................... 29

Figure 9. Grade 6 Mathematics Prototype Item (http://ccsstoolbox.agilemind.com) ......................... 30

Figure 10. The field trip item ........................................................................................................... 45

Figure 11. The growth medium item  ........................................................................................................ 49
Chapter 1: Introduction

Over the past ten years or so we have seen rapid growth in the use of computer-based testing by schools and school districts across the country. Commercial testing organizations, state education departments, and school districts are investing substantial resources in an effort to create valid computer based assessments and data management systems to help teachers, principals, and policymakers track student achievement and identify promising instructional strategies for targeted examinees (Sawchuck, 2009).

With the advent of the newly developed Common Core State Standards (CCSS; http://www.corestandards.org/) and the Next Generation Science Standards (NGSS; http://www.nextgenscience.org/), major shifts are expected in curricula, instruction, and assessment design to ensure that examinees meet these rising performance expectations. Technology-enhanced instructional materials and assessments are believed by many to be integral to the successful development and implementation of these new math and science learning standards. And, it follows, that innovative assessments, including technology-enhanced items and tasks, will be needed to meet the challenges of developing valid and reliable assessments in a world of computer-based testing, some of which may be situated, for example, in so-called ‘epistemic’ games, or games for learning, and other future oriented assessment frameworks and paradigms.

Researchers and test publishers are investing significant resources into developing innovative technology-enhanced items and tasks (TEIs) for use in large-scale, high-stakes math and science tests. The technology available to construct novel item types, it is presumed, will allow for more robust and deeper measures of complex, “hard-to-measure” educationally relevant constructs (e.g., mathematical reasoning, computational thinking, reading
comprehension, etc.). Enhancements in construct representation, item and task formats, and a broader array of examinee response options may, together, enhance our ability to tap hard to measure constructs (Scalise, 2012). Research has long documented that traditional multiple-choice items are not very helpful for validly measuring high-level knowledge and understanding (Archbald & Newmann, 1988; Birenbaum & Tatsuoka, 1987; Lane, 2004; Darling-Hammond & Lieberman, 1992). Often these traditional item types lack the connective tissue, the conceptual assessment framework if you will, between the demands of the assessment tasks, the intended learning objectives, and the instructional practices. High quality educational assessments, in contrast, attend purposefully to the interrelationship of cognition (how examinees process information), observation (the demand characteristics of the test items), and the inferences about student learning flowing from test scores (Assessment Triangle; Pellegrino, Chudowsky, & Glaser, 2001).

Many of the TEIs currently under development (e.g., in the National Assessment for Educational Progress, the Smarter Balanced Assessment Consortium, and the Partnership for Assessment of Readiness for College and Careers), or in use here and there in a variety of operational tests, offer multiple-choice single-answer items and fill-in-the-blank test items, as well as “novel” item types, e.g., puzzle-like questions with multiple correct answers, partially correct answers, and multiple incorrect answers. For example, WestEd designed the simulated-based science program SimScientists (Quellmalz, DeBarger, Haertel, Schank, Buckley, et al., 2008) to support instructional practice in middle school science. SimScientists provides activities and embedded formative assessments in life science, physical science, and earth science. Figure 1 is an example of a life-science activity that requires examinees to draw arrows to mimic the relationship among different organisms.
Alternatively, ETS is in the midst of a long term initiative to develop cognitively based assessment of, for, and as learning (CBAL; Bennett & Gitomer, 2009). Assessments developed under this initiative are intended to be comprehensive in that they not only assess what student know and are able to do, but also allow teachers to diagnose where students’ are struggling so they can adjust their instruction accordingly. Also invaluable to their comprehensive assessment systems, is students and teachers buy-in to the learning opportunities provided by the assessment activities. Figure 2 is an example of a CBAL math item where examinees answer a brief constructed response and provide input from a graphical keypad.
Tasks such as these are presumed to be more meaningful (and valid) indicators of examinees’ knowledge, skills and abilities (KSAs). The reasoning is straightforward and obvious; because TEIs can be designed to represent tasks with greater cognitive complexity, they ought to require the examinee to use deeper and richer focal KSAs that go well beyond simple recognition or recall tasks that are often found in traditional paper and pencil tests comprised of standard multiple-choice items.

For TEIs to be useful, however, they, and the technology-enhanced assessments in which they are embedded, must be well aligned with curricula and, ultimately, integrated with instruction. Performance on TEIs, and the summary test scores they produce, ought to communicate effectively to a host of stakeholders—including students, their parents, teachers and other educators. Simply designing TEIs to be visually attractive and engaging to examinees does not assure their validity from a psychometric or measurement perspective. What TEIs gain in engagement, they lack in efficiency. Although this solves the problem of an examinee’s
inattentiveness that may lead to carelessly select their response, it raises new questions, for example, of dimensionality and construct representation. Item response theory assumes unidimensionality; that a single score on a test implies a single dimension (Messick, 1989). With TEI’s often being activity- or task-based, it becomes more difficult to argue that they are not in fact multi-dimensional. Multi-dimensionality of tasks may even be necessary to avoid construct under-representation given the extensive amount of time it takes to complete many of the tasks.

Also, there is no assurance that the TEIs would necessarily provide the rich forms of information teachers, examinees, and educational policymakers need to improve student learning and teaching more generally. Pellegrino et al. (2001), put a fine point on this problem in the concluding chapter of their seminal book *Knowing What Students Know: The Science and Design of Educational Assessment.*

The principles and practices of educational assessment have changed over the last century, but not sufficiently to keep pace with the substantial developments that have accrued in the understanding of learning and its measurement. It is time to harness the scientific knowledge of cognition and measurement to guide the principles and practices of educational assessment. (p. 313)

Currently, many TEIs are developed one-at-a-time for a specific skill or target construct, or they are developed for use within proprietary systems such as SimScientists mentioned above, thus limiting their portability and reuse. We see this, for instance, in the 2009 NAEP (National Assessment for Educational Progress) science assessment interactive computer tasks (National Center for Education Statistics, 2012). These activity-based tasks were created at the 4th, 8th and 12th grade level and measured narrow constructs at a fine grained level. An example of this is the 12th grade energy transfer task. Before beginning the activity, examinees are informed that
they will be investigating which metal would be best for making the bottom of a cooking pan (Figure 3). They are then given instructions for running the simulation to help understand “specific heat capacity”, that is, the amount of thermal energy required to raise one gram of that substance by one degree Celsius.

Figure 3. Energy Transfer NAEP science activity part 1 (NCES, 2012).
After describing the purpose of the activity, examinees are led to a new screen (Figure 4) where they are instructed to use slider bars to alter the temperature and mass of both the metal and water.

Figure 4. Energy Transfer NAEP science activity part 2 (NCES, 2012).
When they click *Run Simulation* a new screen appears with a graphical representation of their data (Figure 5). Examinees are able to rerun the simulation as many times as necessary to understand how the mass and temperature of a metal affects the temperature of water.

*Figure 5.* Energy Transfer NAEP science activity part 3 (NCES, 2012).
After the examinee has finished running their simulations, they are then asked a series of questions based on the data they collected through the various simulations (Figure 6).

![Energy Transfer NAEP science activity part 4 (NCES, 2012).](image)

Figure 6. Energy Transfer NAEP science activity part 4 (NCES, 2012).

While highly interactive and engaging, this kind of science assessment item took the NAEP test developers several years to create and would most likely be difficult to incorporate into their main operational science assessments given the extended amount of time it takes examinees to complete (NAEP interactive tasks are 20-40 minutes in length). We see this, too, in other assessment development efforts such as the SimScientists (an example of which was presented earlier in Figure 2) project sponsored by WestEd (Quellmalz, et al, 2008) which attempted to measure high-level constructs in engaging ways.

Such assessment systems are expensive and focus extensively on specific content (and often very narrow KSAs) curriculum areas. Most processes used to design and develop TEIs
employ a story-board approach, something akin to developing a movie or a play, in which each step that the student or examinee is expected to make when performing the task is presented as a discrete frame in the storyboard. While this process has yielded several interesting examples of TEIs, it is cumbersome, expensive, inefficient, and often results in writing very specific computer programs. And because this approach focuses on each individual TEI as a discrete object, developers often fail to recognize construct commonalities across tasks. Thus, by focusing discretely on each TEI, developers may miss opportunities to create common classes of task attributes or online tools that can be applied across TEIs. The end result is that most TEIs become “one-off” prototypes that may not lend themselves to the psychometric issues of uni- or multi-dimensional scaling, or to the measurement challenges of reducing construct-irrelevant variance. Nor does the current development approach recognize or capture concerns of cognitive complexity or the cognitive load demands across a variety of TEIs.

Innovations in computer based testing necessitate research to arise and meet these challenges in assessment. Many of these TEIs are not only new and not well understood and/or validated, but they have often been created by groups with little or no background in educational measurement, psychometrics, and cognitive sciences. Although some research has been conducted on different item types such as ordering/ranking, multiple choice item types, generally speaking, are the only item types for which there is a substantial research base in support of their design and use (Haladyna & Rodriguez, 2013). Haladyna and Rodriguez (2013) in their most recent item writing guidelines, do not include guidelines for computer-based formats, but instead highlights the critical need for studying the properties of items and responses to these items. It is essential that through research, we probe the optimal way to design varieties of item types that include the integration of technology.
In this dissertation, I begin by describing the context and reviewing the theoretical and empirical scholarship surrounding the design of large-scale educational assessments. This is followed by the hypotheses tested in this research initiative. I then describe in detail the study design and the analytic methods used to systematically investigate the measurement properties of a subset of TEIs. The findings are then organized in the form of two studies. The first study of which used multi-level modelling to present evidence regarding the use of computer interfaces in educational assessment design and the second study which highlights the applicability of latent class analyses as a useful tool for diagnosing potential design issues. Lastly, I discuss the importance and implications of both studies, their relative limitations, and suggestions for future research.

**Chapter 2: Literature Review**

**Evolution of Technology in Educational Assessments**

In the 1960’s and 1970’s, as computers became increasingly more affordable and readily available, our theories and assessments began their rapid evolution. Computers spurred the cognitive revolution completely changing our view of individuals being a product of behavior to individuals as information processing systems. Around the same time, psychometricians were able to implement item response theory and computerized adaptive testing (Lord, 1971, 1977, 1980; Lord & Novick, 1968, Wright & Stone, 1979; Weiss, 1982, 1983). Unlike its predecessor, item response theory is not sample dependent and provides the ability to examine measurement efficiency of an item at different ability levels. The application of item response theory also contributed to the development of computer adaptive tests. Computer adaptive tests were a pivotal advancement and improvement in testing as we were now able to administer shorter tests with enhanced measurement precision. Additionally, test items were more secure and thus could
be reused for a longer period of time. Lastly, as opposed to paper and pencil exams which must
go through the timely process of being scored, computerized adaptive tests provides scores and
reports immediately.

As technology advanced, other improvements in testing were developed such as
automatic item generation and scoring. Automatic item generation significantly improved the
item development process by decreasing costs and time associated with creating an item bank
while minimizing exposure and improving ability estimates (Gierl and Haladyna, 2013; Wainer,
2002). Traditional item development methods using subject matter experts tends to be incredibly
time consuming and costly. Rudner (2010) estimated that it can cost from $1,500 to $2,000 per
operational item!

Automaticity of processes similarly enhanced the process of scoring by providing the
ability to reduce the time associated with scoring essays, provide consistency in scoring, and
increase efficiency in terms of time and cost (Shermis & Burstein, 2013). There have been
doubts though, in the case of scoring essays, that computers cannot make discerning judgments
about the true qualities of writing given their lack of cognitive processes present in human raters.
But, the cost of human raters appears to outweigh the benefits that construct response items
actually provide in an assessment. Past research has produced evidence that constructed
response items do not improve conditional standard errors or estimated proficiency when added
to a multiple choice test (Lukhele, Thissen & Wainer, 1994). But, with the increased efficiency
provided by automated scoring systems, there is opportunity to decrease cost associated with
scoring and increase their potential use.

More recently, through TEIs, we have been afforded the capability to generate more
complex items that can measure, with greater depth and specificity, the level of knowledge,
skills, and abilities. Research has long documented that traditional multiple-choice items are not especially helpful for validly measuring high-level knowledge and understanding (Archbald & Newmann, 1988; Birenbaum & Tatsuoka, 1987; Lane, 2004; Darling-Hammond & Lieberman, 1992). On the other hand, by leveraging the power of the computer, TEIs can present tasks that advances the level of complexity and demands to extended responses that are amenable to automated scoring algorithms embedded in the computer’s operating system (Archbald & Newmann, 1988). And thus, enriching our measurement of constructs that tend to be problematic due to their intricacy (Scalise, 2012). Often, this has been observed in the development of simulations, e.g., those used for medical assessments, which evaluates an examinee’s reasoning throughout the steps taken to successfully diagnose a patient.

Game-based assessment is an exciting new development that is not only engaging, but has the potential to produce various data on interactions thus contributing to the information gathered on what examinees know and are able to do. Furthermore, examinees are afforded opportunities to acquire new complex skills by learning through interactive simulations. Embedding assessments into games, for example, allows us to detect proficiencies while at the same time examinees are engaged and able to learn (Kopfer, Osterweil, & Salen, 2009; Wang, Shute, & Moore, 2015).

**Contemporary Views on Assessment Design**

With the exponential growth in technology and its influence on assessment and item development, advances are needed in frameworks for assessment design to assure the validity, replicability, and generalizability for various computer-based items and tests. Current guiding interpretive frameworks focus on a construct-centered approach to validation. This approach, as citing Samuel Messick (1993),
…would begin by asking what complex of knowledge, skills, or other attributes should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or are otherwise valued by society. Next, what behaviors or performances should reveal those constructs, and what tasks or situations should elicit those behaviors? Thus, the nature of the construct guides the selection or construction of relevant tasks as well as the rational development of construct-based scoring criteria and rubrics. (p. 17)

Messick’s (1993) ideas of a construct centered approach, was later used to develop today’s predominant framework, Evidenced Centered Design (ECD) (Mislevy, Almond, & Lukas, 2003; Mislevy & Haertel, 2006; Haertel, Vendlinski, Rutstein, DeBarger, Cheng, et al, 2016). With ECD, assessments are viewed in a validity framework, that is, an argument that presents claims and evidence that support the interpretation and use of scores (see Figure 7). The argument articulately conveys potential claims about what constitutes student proficiency in a domain (student model); identifying the kinds of things examinees might say or do that would constitute evidence about these proficiencies (evidence model); and identifying the kinds of situations or tasks that might produce this evidence (task models).
Figure 7. Structure for assessment arguments (Mislevy, 2008).

Five layers are used to describe the form of ECD: domain analysis, domain modeling, conceptual assessment, assessment implementation, and assessment delivery. During the first layer, domain modeling, the construct of interest and features of mastery are gathered and structured into a clearly defined argument. For relative guidance, design patterns are developed, centered on the relative applicable knowledge, skills, or abilities, as tools to support design choices. Intentionally, design patterns are broad, with content and level of detail included in addition to those features that are essential in a problem setting if it is to evoke evidence. They are not technical, but rather a tool that can facilitate the acquisition of evidence about those knowledge, skills or abilities. Given the reusability and strengthening of validity through the inclusion of background information and rationales, design patterns are essential to the process of designing tasks with the ECD framework.
The blueprint for an assessment is laid out in the conceptual framework where technicalities are specified with, as referenced above, a student model, task model, and evidence model. The student model portrays what variables being measured are reflective of student proficiency, whereas the task model in essence describes the testing environment. Lastly, the third model within the conceptual framework, the evidence model, includes evaluation and measurement, bridging the student model and task model.

The fourth and fifth layers of ECD include assessment implementation and assessment delivery. Their concerned with preparing operational elements and the process from examinees participating in the assessment to reports that are produced.

Although a highly efficacious framework when implemented correctly, advances are crucial to meet the increasing demands of complex, technology-based tests. Luecht (2013), recommends Assessment Engineering (AE), a fairly new approach, based on engineering principles which, unlike ECD, provides the ability to “manufacture” items. In contrast to ECD, AE, although also a construct-centered approach, requires constructs to be articulated on a score scale with proficiency claims at the beginning of the process. Task models, which include several different yet related items, are crafted with an intended position on the proficiency scale, and in a manner that makes them reliably replicable. Lastly, a calibration system is used for quality control to maintain the statistical and interpretive integrity of the score scale. AE lends itself, unlike other frameworks, to the specification of constructs that are not only highly detailed, but also includes the hierarchical progression through mastery of a construct.

In AE, traditional content blueprints are replaced by task models aligned on a construct map. These models include objects and their properties, nature of relationship among objects and their persistence, and functional classes that represent the action required on the objects and
any special conditions. Alignment in task models to the construct map are in the order of difficulty and requires empirical confirmation. Following, multiple templates that behave similarly in their psychometric properties are produced from the task models and used to design items.

Item design is structured through the use of templates to promote “manufacturing”. There are specification for the control of the aesthetics and format of the items and control in how the item is scored. This helps ensure reusability and scalability as well as minimizes item writers from contaminating the item to misrepresent the construct and relative difficulty.

**Current Understanding of Item Design**

Although ECD and AE provide invaluable frameworks, assessment design need still be supported with research aimed at understanding the impact of an item’s design features on the validity of score interpretation. Currently, the item type we have the clearest understanding of is the traditional multiple choice, given its extant amount of research. It has been argued that approximately 90 years of research and development activities on multiple choice item formats supports the claims that they efficiently and effectively measure cognitive ability. They can measure higher order cognitive abilities, a large sample of the cognitive domain (Haladyna & Downing, 2004; Messick, 1989), and are stem equivalent to constructed response items (Rodriguez, 2003). Furthermore, extensive empirical research has been conducted to assess particulars of the item design that may contribute to performance on the item, particularly the number of response options. A meta-analysis of these studies concluded that three options are optimal (Rodriguez, 2005).

Furthermore, important to stakeholders is the fact that these items are cost effective. In examination of an AP Chemistry exam, Wainer and Thissen (1993) found that the constructed
response portion must be three hours long to reach a .92 reliability whereas the multiple choice section in chemistry only required 75 minutes. Not only did they find that constructed response sections require an inordinate amount of time to obtain optimal reliability, but also the cost to obtain that level of reliability is exponentially higher than for multiple choice sections. As the desired reliability increases ($30 for a reliability of .92), the cost of scoring constructed response sections increases exponentially. All while the cost of multiple choice tests remains at a fraction of the cost ($0.01) regardless of the reliability. Unfortunately, this even varies across different subjects. For example, the now discontinued AP Music test required 26 hours of testing time to reach a reliability of .92 and cost $100 to score!

Regardless of the considerable amount of research, score interpretations from multiple choice items can still lack adequate validity. But, Haladyna, Downing, and Rodriguez (2002) maintain that this is not a result of the type of item itself, but rather poor item writing. Experts agree that effective item writing requires not only expertise in the subject matter, but also training in the art of producing effective test items. Currently, there are 31 consensus principles of effective multiple-choice item writing to help guide item writers to lessen the chance of poor items being written (Haladyna, 2002 review paper).

From an IRT perspective, multiple choice items provide about twice as much information in the same amount of time than constructed response. In addition, constructed response items, when added to a multiple choice test, does not improve conditional standard error of estimated proficiency (Lukhele, et al, 1994). Next generation assessments purport to include various item types beyond multiple choice and constructed response. Because item types differentially effect reliability, testing time, and cost of scoring, IRT weighting will become extremely pertinent.
Despite the evidence gathered supporting multiple choice item types, measurement experts attest that more complex item types can improve measurement of certain constructs, specifically higher-level constructs (Bennett, 1999; Harlen and Crick, 2003; Huff and Sireci, 2001; Jodoin, 2003; McFarlane, Williams, and Bonnett, 2000; Sireci & Zenisky, 2006; Zenisky & Sireci, 2002). Although one specific complex item type has yet to be as thoroughly investigated as the traditional multiple choice format, research suggests that many of these item types provide better measurement of certain constructs, specifically higher-level constructs (Bennett, 1999; Harlen and Crick, 2003; Huff and Sireci, 2001; Jodoin, 2003; McFarlane, Williams, and Bonnett, 2000; Sireci & Zenisky, 2006; Zenisky & Sireci, 2002). Evidence suggests that drag-and-drop and create-a-tree item formats provide more information about examinees’ skills (Jodoin, 2003). Masterman and Sharples (2002) found that items asking examinees to depict their understanding of a concept using a diagram provided better, richer diagnostic information about examinees’ knowledge structures. Others have reported that, although these complex item types often take more time to complete, they do indeed provide more information about examinees’ cognitive abilities (Martinez, 1999). In a series of studies conducted by Randy Bennett and his colleagues at ETS (Bennett & Rock, 1995; Enright, Rock, & Bennett, 1998), item-types asking examinees to formulate hypotheses were observed to be useful for assessing dimensions of scientific reasoning on the Graduate Record Exam (the GRE). More recently, researchers learned that examinees often use weak reasoning processes based on option elimination when presented with multiple-choice items that require examinees to select the correct response option. But, when presented with a more complex task, e.g., constructing an appropriate scatter plot given a set of data, the task appears to be more difficult, demanding
greater knowledge, skills, and cognitive load (Dolan, Goodman, Strain-Seymour, Adams & Sethuraman, 2011).

The Role of Cognitive Psychology in Assessment Design

In a recent critique of the next generation assessments in math (i.e., *Smarter Balanced*), Rasmussen (2015) observed that many aspects of the technology “enhancements” can be expected to do more harm than good. Some mathematics problems, he found, included technology “enhancements” that were not relevant to the task at hand, and thus unnecessary. In addition, he found inconsistency in the design of the interfaces used to answer an item, for example, alternating between requiring the use of a graphical keypad on the screen versus requiring the use of the actual keyboard. Not only was there this inconsistency, which would amount to confusion among test takers, but the graphical keypad was poorly designed. Calculators, as we all know them, automatically correct for some errors. For example, if an examinee were to mistype a number by inadvertently adding two decimal points (e.g., 45..2), the graphical keypad on the screen would not automatically correct for this mistake as would a calculator.

Rasmussen (2015) also exhibited trepidation about ‘drag-and-drop’ interfaces with ‘snap-to’ behaviors. When an examinee drags an object to, for instance, a number line and attempts to place the object between two whole numbers, the object would unexpectedly ‘snap-to’ the whole number on either side. If the correct answer was 4 and an examinee attempted to place the object in the middle of 3 and 4, let’s say to represent 3.5, the object may, on its own volition, ‘snap-to’ 3 or 4. Depending on the parameters set for this interface, an examinee’s answer may well be right or wrong.
Computer interfaces, therefore, may introduce construct irrelevant variance, as an examinee must not only demonstrate their knowledge of mathematics, but also on how to use the computer interface. Cognitive load theory (Miller, 1956) is a promising framework that can be applied to computer-based item design to mitigate the effects of the usability of the interface. Otherwise known as working memory capacity (Baddeley, 1983) or cognitive energy (Miyake, Just, & Carpenter, 1994), cognitive load is essential to an examinee’s performance on a test item as it directly affects their ability to process and act on information presented to them. If cognitive load is depleted, examinees lack the capacity in their working memory to hold information presented to them and are thus less likely to respond accurately to an item. In other words, if examinees cognitive load is depleted from expending mental energy to understand how to use the computer interface, they may not have sufficient working memory available to attend to actually solving the mathematics items, thus, increasing the probability of answering the item incorrectly.

Cognitive load is theorized to be composed of three different types of load: intrinsic, extraneous, and germane. Intrinsic and extraneous cognitive load both involve elements a student must process within the materials they are learning. The elements can either be essential or irrelevant to the cognitive processes or knowledge to be acquired. Those that are essential are referred to as imposing intrinsic cognitive load. If a task contains a greater number of interacting essential elements a student must process simultaneously, then more intrinsic load imposed. But, the relative impact of those essential elements on an individual’s intrinsic cognitive load depends on their relative experience. This is because, a examinees with more expertise in the subject or task at hand, would not need to simultaneously process all of the essential elements due to already having an active schema.
Research on cognitive load effects (Table 1) has identified three different cognitive load effects primarily identifying with sources being intrinsic: the goal free effect, element interactivity effect, and the isolated/interacting element effect. Computer administered item design may particularly benefit from isolating elements in an item that is very complex with several different interacting elements. For example, breaking up a very complex item into several parts instead of presenting it all as one question would enable examinees to only need to process parts of the information at a time, instead of all of the essential elements simultaneously.
Table 1. Cognitive Load Effects (Sweller, 2010).

<table>
<thead>
<tr>
<th>Cognitive Load Effect</th>
<th>Description</th>
<th>Primary Cognitive Load Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Worked-Example</td>
<td>Studying worked examples results in better performance on subsequent tests of problem solving than solving the equivalent problems (Renkl, 2005)</td>
<td>Extraneous</td>
</tr>
<tr>
<td>Completion</td>
<td>Requiring learners to complete partially solved problems can be just as effective as worked examples (Paas &amp; van Merrienboer, 1994)</td>
<td>Extraneous</td>
</tr>
<tr>
<td>Split-Attention</td>
<td>Multiple sources of information that are unintelligible in isolation result in less learning when they are presented in split-attention as opposed to integrated format (Ayres &amp; Sweller, 2005)</td>
<td>Extraneous</td>
</tr>
<tr>
<td>Modality</td>
<td>Multiple sources of information that are unintelligible in isolation result in less learning when they are presented in single-modality as opposed to dual-modality format (Low &amp; Sweller, 2005)</td>
<td>Extraneous</td>
</tr>
<tr>
<td>Redundancy</td>
<td>The presence of sources of information that do not contribute to schema acquisition or automation interfere with learning (Sweller, 2005)</td>
<td>Extraneous</td>
</tr>
<tr>
<td>Guidance Fading</td>
<td>With increasing expertise, learners should be presented worked examples followed by completion problems and then full problems rather than worked examples alone (Renkl, 2005)</td>
<td>Extraneous</td>
</tr>
<tr>
<td>Goal-Free</td>
<td>Problem presented in goal free form enhance learning compared with conventional problems (Paas, Camp, &amp; Rikers, 2001)</td>
<td>Intrinsic</td>
</tr>
<tr>
<td>Element Interactivity</td>
<td>Cognitive load effects are only obtainable using high rather than low element interactivity material (Sweller, 1994).</td>
<td>Intrinsic</td>
</tr>
<tr>
<td>Isolated/Interacting Elements</td>
<td>Learning is enhanced if very high element interactivity material is first presented as isolated elements followed by interacting elements versions rather than as interacting elements form initially (Pollock, Chandler &amp; Sweller, 2001).</td>
<td>Intrinsic</td>
</tr>
<tr>
<td>Variable Examples</td>
<td>Examples with variable surface features enhance learning compared with examples with similar features (Paas &amp; van Merrienboer, 1994).</td>
<td>Germane</td>
</tr>
<tr>
<td>Imagination</td>
<td>Imagining procedures or concepts enhance learning compared with studying materials (Leaby &amp; Sweller, 2001).</td>
<td>Germane</td>
</tr>
</tbody>
</table>

Those elements that are irrelevant to the cognitive process or knowledge to be acquired impose extraneous cognitive load. Similarly to intrinsic cognitive load, extraneous cognitive load is impacted by the number of irrelevant elements in a given task. When an overwhelming
amount of extraneous load is present, it distracts examinees from being able to process the relevant information to be learned. Compared to intrinsic cognitive load, research on extraneous cognitive load has been more productive in identifying different effects (worked-example, completion, split-attention, modality, redundancy, and guidance fading) (described in Table 1). Furthermore, many of these effects have been applied to multi-media learning with the following guiding principles: (1) coherence principle: delete extraneous material from multimedia instruction; (2) redundancy principle: exclude redundant on-screen text from narrated animations; (3) signaling principle: incorporate signals in the narration, such as outlines, heading, and pointer words; (4) temporal contiguity principle: present corresponding segments of animation and narration concurrently; (5) spatial contiguity principle: place on-screen text near corresponding elements in the screen; and (6) segmenting principle: break narrated animations or narrated videos into learner controlled segments (Mayer & Moreno, 2010).

The third aspect of cognitive load is germane cognitive load. When a student is learning new materials, schemas—techniques for organizing knowledge in memory—are being acquired to allow for automation of different processes. When a student is undergoing schema acquisition, germane cognitive load is imposed. For example, suppose a student is learning how to solve the following mathematical problem: 3(2+6). Examinees will develop schemas that allow them to automatically recognize that they must first solve for what is within the parentheses. Because of the novelty of many of the TEIs, practice items may help reduce the germane cognitive load.

Unlike design frameworks for computer-based assessments and TEIs, the cognitive load theory framework is common knowledge among computer interaction experts (Rodgers, 2004; Van Nimwege, Van Oostendorp, Burgos, & Koper, 2006). One of their central tenets for
development is to reduce the cognitive load for users as much as possible (Mandel, 1997; Preece, Sharp, & Rogers, 2002). Human computer interaction research has provided evidence that cognitive load theory is a useful framework for guiding work on the design of learning environments and educational computer systems, e.g., learning management systems (Chalmers, 2003; Sawicka, Kopainsky, & Gonzalez, 2008). Their investigations have revealed that cognitive load theory has the power to improve the usability by divesting users from the cumbersome act of maintaining information in their brain from one screen to the other. High working memory load is resolved by using generic commands, keeping displays simple and clear, offering functionalities only in the context in which they are needed, and training users when complex interactions are required (Mandel, 1997; Nielsen, 1994; Shneiderman, 1998; Van Nimwegan et al, 2006).

Surprisingly, this has not been within the predominant framework driving the design of computer-based assessments despite a high level of apprehension surrounding the usability of the tool or application (Chalmers, 2003; Sharp, Roger, & Preece, 2007). Only very recently have these ideas emerged in research. Recently, Gillmor, Poggio and Embretson (2015) investigated experimentally the applicability of CLT to improving the validity of items. Examinees from three different schools were randomly assigned to an experimental or control condition and given a paper-based mathematics exam. A traditional test made from commercially available items was given to the control groups whereas the experimental group received the same items, but modified for reduced cognitive load. The modifications included reducing word count and simplifying language, using a diagram to represent spatial information, focusing attention with signals, eliminating extraneous visuals and text, sequencing of the question and answer options, placing text near corresponding features on figures, and using simpler numbers. Results revealed
that those in the experimental group performed better than those in the control group, suggesting that reducing cognitive load features improves the validity of the measurement instruments. Specifically, the modifications that were most effective included focusing examinees’ attention with signals (e.g., underlining important words), eliminating extraneous visuals and text from the test items, and placing text near corresponding features on figures.

In an examination of modifications for alternate assessments, Kettler et al (2011) found that removing a response option and simplifying language were effective in reducing extraneous load. In addition, graphics included in reading items that were intended to increase interest may increase cognitive demand.

Hess, Johnston & Lipner (2013) extended this line of research examining the effect item and examinee characteristics have on response time. Data were gathered from a computer-based medical certification exam where examinees were administered a section of 60 multiple choice items. A response times-within-person growth modeling was conducted with 60-time points relating to the individual items. Level-1 included the following item characteristics: item format, word count, classical item difficulty, and presence of graph or picture. Level-2 included student characteristics such as gender, age, first time test taker, and ability. Their results suggest that item characteristics cumulatively explained 41% of the total within-person variation in log response time. More specifically, item difficulty, word count, and graphics were associated with greater response times. When taking examinee characteristics into account, women and those with higher ability, in particular, tended to spend more time on items that contained graphics and a greater number of words.

Despite the limited amount of research applying cognitive load theory to assessment design, measurement experts agree that assessments cannot be designed without consideration of
a cognitive model. This idea, though, typically extends to specifications for the types of observations needed to provide evidence, and how to appropriately interpret the evidence gathered. But, as aforementioned, assessment should be viewed as a triangle with all three vertices of the triangle in synchrony, connecting to each other in meaningful ways (Pellegrino et al., 2001). Despite the strong agreement among measurement experts, many large scale assessments today collect a multitude of observations, but have no real cognitive model underlying what those observations are meant to portray (Leighton & Geirl, 2011; Ferrara & DeMauro, 2006).

Given the rapid rate at which the design of assessments are currently changing to include higher-level tasks and computer interfaces, it is disconcerting that a cognitive model is not readily applied to all aspects of next generation assessments design. Pellegrino, Wilson, Koenig, and Beatty (2014) recommend that test developers take advantage of innovations in assessment design (which also includes innovations in user interfaces), scoring, and reporting to create items and tasks that tap the multi-dimensional aspects of learning demanded by the new learning standards.

**Hypotheses**

While providing many advancements in testing, technology is growing at such a rapid rate that research into its effect on testing lags behind. More purposeful attention to cognitive load theory as a tenet of item develop is needed. Although the reduction of cognitive load is a primary goal in human computer interaction research, consideration of its application to computer-based testing is lacking. This is important as the usability aspect of the technological tools or applications may affect how an examinee responds to an item; poor design of usability will likely increase an examinee’s likelihood of producing an incorrect answer despite the fact
that they have the sufficient knowledge, skills, and abilities. The studies described below demonstrate the applications of multi-level modeling and latent class modeling to the evaluation of item development.

In the first study I used multi-level modeling to assess the effect of item characteristics on examinees’ relative performance. I hypothesized that item level characteristics, namely response format, would significantly contribute to the amount of variance explained by item characteristics over and above student characteristics. Specifically, that examinees would more often correctly answer multiple-choice-single-answer than multiple-choice-multiple-answer, drag-and-drop, and fill-in-the-blank.

In study two, I used two exemplar items to show how important information about variation in response to novel-items can be gleaned by examining examinees’ response patterns as opposed to scores. During the prototyping phase of item development, further fine grained analyses may give clues into potential problems with the item or its corresponding scoring rubrics. Data concerning examinees’ actions—produced through latent class analyses—can be used as evidence in validity investigations of items to provide additional information in conjunction with evidence provided by item response theory.

Chapter 4. Study Design

Data

Both studies utilized data from an earlier, stand alone, pilot study of mathematics prototype items (Everson, Verkuilen, Thomas, and Racanello, 2012). Prototype items were designed by mathematics educators for grade 3, 4, 6, 7 and 8/9 (Algebra). These items were bundled into tasks in order that a cluster of items provided evidence for the common core state standards measured. About 59 prototype tasks were developed which together included a total of
146 individual items. Design for these tasks was based on three evidence-centered design task models which included specifications in the form of sub-claims and mathematical practices (Appendix A). In addition to each task being aligned to standards and mathematical practices as defined by the CCSS (Appendix B). All tasks were computer-based and used both multiple choice and innovative item designs such as ‘drag-and-drop’ puzzle like tasks (see Figure 8) or used simple simulations to explore relationships (see Figure 9).

Julia is planting flowers. She wants to cover $\frac{3}{4}$ of the garden with flowers.

Drag a tile onto Julia’s garden that will finish covering $\frac{3}{4}$ of her garden with flowers.

Figure 8. Grade 3 Mathematics Prototype Item (http://ccsstoolbox.agilemind.com)
A field study was then conducted to test the feasibility of designing and delivering innovative, next generation math tasks. Could they be designed in ways that take advantage of computer technology? Would they illustrate the math standards with fidelity, and be machine scored? Would they be too easy or too challenging for examinees? It was also of interest whether the prototypes could be built in a way that would measure examinees’ mathematical knowledge and reasoning skills as described in the common core state standards. Could they assess both the overarching habits of mind represented in the Standards for Mathematical Practices, and measure growth in conceptual understanding expressed in the grade-by-grade
progression of the math standards? If so, the prototypes would serve well as exemplars and templates for the next development step—creating larger and durable pools of next generation math assessment tasks. These were the questions and issues behind the prototyping effort.

**Sample.** Examinees in grades 3, 4, 6, 7 and 9 were drawn from two large public school districts from two different states (herein referred to as school district A and B for purposes of anonymity). Given the complexity of mounting the pilot study in public school settings, and the myriad logistical issues of field-based studies generally, a number of assumptions were made as the pilot study was designed the pilot study. For example, the following design constraints influenced planning and administration of the field study.

1. Limited testing time, roughly no more than 50 to 60 minutes of actual test administration time. This required administration of relatively small sets of prototypes—no more than six or seven tasks per bundle, and required the creation of two non-overlapping bundles per grade level to ensure exposure of all the prototypes.

2. The tasks were to be administered using both computer-based and paper-pencil response formats, i.e., computer-based testing would be limited to short, binary response-type items or tasks while the extended response portions of the tasks were completed offline using paper-and-pencil test booklets.

3. All binary response-type items were computer scored, while the more complex extended response tasks, administered offline, were scored by recruited and trained math educators.

**School district A pilot study site.** School district A was one of the larger public school district in state 1. It was comprised of 173 schools and enrolls more than 96,500
examinees from kindergarten through high school. A large proportion of Hispanic or
Latino examinees were enrolled, many of whom were non-native speakers of English,
and many were eligible for federally sponsored free and reduced lunch programs.
Recruiting included 1,195 examinees from this school district. The data suggested that
about 30-35% of the examinees were classified as less than proficient in reading and/or
math on the states 2011 standards-based assessments.

**School district B pilot study site.** School district B was one of the larger public
school districts in state 2. It encompassed more than 305 schools that served more than
193,000 examinees. There was a significant proportion of Hispanic or Latino examinees
(about 20%) and an equally large proportion of African-American examinees (18% or
so). Many (about one-quarter) of the examinees were non-native speakers of English, and
many were eligible for federally sponsored free and reduced lunch programs. Recruiting
included 926 examinees from this school district. Data from the state’s testing program,
suggested that about 30-40% of the examinees we sampled were classified as less than
proficient in reading and/or math.

**Procedures for Conducting Field Study**

Participating examinees at each grade level were administered one of two bundles
of prototype tasks (referred to as Forms A and B). Each bundle or test form was
comprised of six or seven prototype tasks, depending on the grade level, with most test
forms containing six prototype tasks.

We attempted whenever feasible to administer the bundles of items as spiraled or
alternating forms. All prototypes were administered via computer, and many of the examinees’
responses were captured and scored by the computer. As we noted earlier, for the prototypes
demanding more complex constructed responses, the examinee was instructed to provide their answers and show their work in the paper-and-pencil test booklets that were provided at the time of testing.

Examinees responses and response times were captured by the computer. Subsequently, the responses were scored and demographics collected. Below is a description of each of the information collected.

**Examinee responses.** The computer captured what response option(s) an examinee selected. For example, if an examinee was presented with an A-D 4-point multiple-choice-single-answer item and select option C as the correct response, the computer recorded a C. For those items that contained multiple possible answers, the computer captured the string of responses. Using the previous example, if it were instead a multiple-choice-multiple-answer item and a student selected options A, B, and C as the correct response the computer recorded A, B, C. This enabled us to capture what exactly the examinees’ responses were before being scored, allowing for the scoring rubric to be compared to the response patterns.

**Scores.** During the item development phase, scoring rules or rubrics were created in conjunction with the item prompts by the item-writers; most of whom were mathematics educators. The scoring rules varied by item with some rubrics assigning binary scores (correct vs. incorrect) and some additional assigning partial credit scores (e.g., 0, 1, or 2 points possible).

**Response time.** The number of seconds examinees’ spent on a “page” was recorded by the computer. Instead of recording the amount of time that lapsed before clicking the correct answer, the computer recorded the time that lapsed before moving to the next item.

**Examinee demographics.** Various demographics were collected from the school district partners. This included gender, race/ethnicity, English language learner designation, special
education designation, and prior state accountability test scores. Given that the pilot was conducted at school districts in two different states, scores were standardized for comparability purposes in analyses.

Participants

As mentioned previously, the 2,121 examinees who participated in this study were recruited from two large school districts in two different states. Those examinees who participated were drawn from grades 3, 4, 6, 7 and 9 (algebra courses). Table 2 shows the number of examinees across the two districts by grade level. More detail on the demographic characteristics of the examinees who participated can be found in Table 3 below.

Table 2. Distribution of Examinees Tested by Grade Level.

<table>
<thead>
<tr>
<th>N of Examinees</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>N of Examinees</td>
<td>2121</td>
<td>100%</td>
</tr>
<tr>
<td>N Per Grade Level</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3rd Grade</td>
<td>355</td>
<td>17%</td>
</tr>
<tr>
<td>4th Grade</td>
<td>480</td>
<td>23%</td>
</tr>
<tr>
<td>6th Grade</td>
<td>442</td>
<td>21%</td>
</tr>
<tr>
<td>7th Grade</td>
<td>387</td>
<td>18%</td>
</tr>
<tr>
<td>9th Grade</td>
<td>457</td>
<td>21%</td>
</tr>
</tbody>
</table>
A broad and diverse group of examinees took part in the study. The pools of examinees were, for the most part, balanced with respect to gender across both sites. However, there were differences in the proportions of Hispanic or Latino examinees and the percentages of African-American examinees. School district A enrolled many more Hispanic or Latino examinees, while School district B had a large proportion of African-American examinees in the study. The sample was relatively balanced with respect to English language literacy levels as well.
Chapter 5. Study 1

In the following study I tested whether the type of response format used (i.e., multiple-choice-single-answer, multiple-choice-multiple-answer, drag-and-drop, and fill-in-the-blank) accounted for a significant amount of the variation in examinees’ performance on an item. Furthermore, analyses were extended to reflect whether the above was mitigated by examinee characteristics (i.e., gender, race/ethnicity, free-reduced lunch, and prior achievement in mathematics and reading.

Method

Multi-level models. Multilevel logistic regression with 2-levels was conducted to assess the impact of the various item formats on examinees’ scores over and above student level characteristics. Multi-level modeling (MLM) is a class of techniques for analyzing data having a hierarchical or nested structure (Raudenbush & Bryk, 2002). For example, this study examined examinees nested within items. Much like students may be assigned by some mechanism to teachers, examinees in the pilot study were assigned to take different sets of items. Using multi-level modeling methods allowed the investigation of differences in item characteristics across items when student characteristics were taken into account while simultaneously allowing me to model the independent effects.

Conventional regression techniques were not appropriate as they would lead to ignoring the nesting of these structures by treating the student as the unit of analysis or ignoring the variation among examinees or among items by treating the item as the unit of analysis. In contrast, the MLM technique allowed me to overcome problems that would be produced by conventional regression techniques such as biased estimates of standard errors and biased
regression weights. MLM allows me to model student level regressions for each item, thus allowing standard errors of means and regression coefficients to be estimated without bias.

**Level-1 student variables.** For level-1 predictors, the recommended procedure for selection of the best subset to be tested in an MLM analysis is through preliminary modeling using linear regression (Raudenbush & Bryk, 2002, p. 257). Based on previous research on variables influencing academic performance, the following variables are expected to be included in the MLM analysis: gender (Maerten-Rivera, Myers, Lee, & Penfield, 2010; Bacharach, Baumeister, & Furr, 2003; Mae & Lynn, 2000), race/ethnicity (Maerten-Rivera et al, 2010; Bacharach et al, 2003; Chapin, 2006), free-reduced lunch, and prior achievement in mathematics and reading scores (i.e., state accountability test scores standardized for comparability).

**Level-2 (item-level) variables.** Item formats were entered as predictor variables in the item-level model. This included (1) multiple-choice-single-answer (MCSA); (2) multiple-choice-multiple-answer (MCMA); (3) drag-and-drop (DND); (4) fill-in-the-blank (FIB).

**Model Testing Sequence.** Several stages of model testing were performed to test carefully this 2-level model. First, fully unconditional models were fit to estimate the variance decomposition given by the equations 1 and 2:

\[
\begin{align*}
\text{Level 1: } Y_{ij} &= \beta_{0j} + e_{ij} \\
\text{Level 2: } \beta_{0j} &= \gamma_{00} + r_{0j}
\end{align*}
\]

This two-level model partitions the total variability in the outcome \(Y_{ij}\) into the following two components:

\[
\begin{align*}
\sigma^2 / (\sigma^2 + \tau_{\pi} + \tau_{\beta}) & \quad (3) \\
\tau_{\pi} / (\sigma^2 + \tau_{\pi} + \tau_{\beta}) & \quad (4)
\end{align*}
\]
Where equation 3 represents the proportion of variance within schools, equation 4 models the proportion of variance among items. Predictors were included based on whether there is a substantial amount of variance to be explained at each level. After determining which predictors to include, the conditional models were tested. At the student-level, item means adjusted for student characteristics were estimated through a logistic regression. Thus, the student-level model was as follows (see Appendix C for further explanation of the coefficients):

\[
\text{Level 1: } Y_{ij} = \beta_{0j} + \beta_{1j}a_{1ij} + \cdots + \beta_{pj}a_{ pij} + e_{ij} \quad (5)
\]

At the item-level, regression methods were used to identify those item formats that account for variation among item means, after adjusting for differences in student characteristics. That is, whether the item-level outcomes (i.e., the coefficients from the student-level model) were predicted by item-level characteristic plus a random item-level error was assessed. The item-level model was as follows (see Appendix C for further explanation of the coefficients):

\[
\text{Level 2: } \beta_{pj} = \gamma_{p0} + \sum_{q=1}^{Q_p} \gamma_{pq} X_{qj} + r_{pj} \quad (6)
\]

Results

Multi-level logistic regression was conducted using STATA/SE 14.0 (StataCorp, 2015) to statistically analyze a data structure where examinees (level-1) were nested within items (level-2). Of specific interest was the relation between examinees' scores (level-1 criterion variable), student characteristics (level-1 predictor variables), and item characteristics (a level-2 predictor variables). Model testing was conducted in four phases: intercept-only model, means-as-outcome model, random-regression coefficients model, and the intercepts- and slopes-as-outcomes model (refer to Appendix D for relevant STATA syntax).

The intercept-only model was performed to assess the amount of variability that exists within and between items. This model revealed an intraclass correlation coefficient of .53. This
suggests that half of the unexplained variance exists between-items and half exists at the student-level. Because variance existed at both levels of the data structure, predictors were individually added at each level to assess whether those predictors explain any of the variance.

The means-as-outcomes model was performed to assess whether examinees’ score on an item was predicted by item characteristics. Therefore, item type was added as level-2 predictor dummy variables (i.e., DND, FIB, and MCMA). The regression coefficients relating the items types to examinees’ scores were not significant: DND ($\beta^1 = -.191, p = .793$), FIB ($\beta = -.602, p = .394$), and MCMA ($\beta = -1.37, p = .112$).

Next, the random-regression coefficients model was performed to assess the average intercept and slope across items and whether these parameters varied from item to item. Thus, examinees’ gender, race/ethnicity, free/reduced lunch status, and prior math and reading scores were included as the predictor variables. The regression coefficient relating examinees’ gender to examinees’ score was positive and statistically significant ($\beta = .182, p = .010$). Male examinees were 1.2 times more likely to answer an item correctly than females. Free reduced lunch status was negative and statistically significant ($\beta = -.241, p = .002$) suggesting those who do not have free/reduced lunch status were more likely to answer an item correctly. The regression coefficient relating examinees’ ethnicity to examinees’ score was positive and statistically significant for Asian/Pacific Islander ($\beta = .569, p = .002$); negative and statistically significant for Hispanic ($\beta = -.301, p < .001$) and American Indian/Alaskan Native ($\beta = -.571, p = .006$). Black, on the other hand, was not statistically significant. Asian/Pacific Islanders were 1.77 times more likely to answer correctly than other race/ethnicity. The regression coefficient

---

1 Note: $\beta$ is being used in the context of the results to refer to the parameter associated with the slope for each model. This was done in order to assure ease of interpretation by knowing that all parameters that are being discussed are slopes.
relating examinees’ prior math scores was positive and statistically significant ($\beta = 2.91$, $p < .001$) whereas the prior reading scores were negative and statistically significant ($\beta = -2.65$, $p < .001$). This suggests that those with higher prior math scores and lower prior reading scores were more likely to answer an item correctly.

Finally, the intercepts- and slopes-as-outcomes model was tested with all predictors tested in the model simultaneously. All regression coefficient relating student characteristics to examinees’ scores continued to be the same as reported above. The cross-level interactions between examinees’ scores and item type were not statistically significant DND ($\beta = -.255$, $p = .741$), FIB ($\beta = -.761$, $p = .313$), and MCMA ($\beta = 1.59$, $p = .082$).

**Summary of Findings**

Results from Study 1 suggest that item type does not explain the variation in student scores. Student characteristics were much more highly related to examinees’ relative performance on an item. This suggests that examinees are familiar enough with different modes of answering items on a computer where the effect of item type on answering correctly was negligible.

**Chapter 6. Study 2**

In the following study I provide evidence through two exemplar items in how examining response patterns, opposed to scores designated by the scoring rules/rubrics, can provide item developers information helpful in flagging potential issues with the item or its corresponding rubric.

**Method**

**Exemplar items.** Although a number of multiple-choice-multiple-answer items were present in the dataset, only two were chosen as exemplars for this study. The exemplars were
chosen as they were representative of two different response patterns results common among all of the multiple-choice-multiple-answer items. The first item is a good example of the tendency for examinees’ to choose one answer that was a correct answer and the consequences of designing a scoring rubric that does not allow partial credit for those correct answers. The second item provides an example of the large number of possible response patterns also inherent in many of the items.

**Latent class analyses.** Latent Class Analyses (LCA; Lanza & Collins, 201X, Lazarsfeld & Henry, 1968) were conducted using MPLUS 6 (Muthen & Muthen, 2007) on two exemplar items to understand further examinees’ responses to multiple-choice-multiple-answer item types (refer to Appendix E for relevant MPLUS syntax). LCA allows for the identification of classes of examinees based on similar response patterns. This is done by producing probabilities that an examinee is a member of some latent class given their item responses and the probability that the examinee will select a response given a latent class.

In the two exemplar multiple-choice-multiple-answer item types, I wanted to understand how examinees typically responded to these types of technology enabled items. Given that examinees can essentially select all responses they believe apply, this allows for a large number of different response patterns. For example, an item that provides 6 options has the potential for 720 different response patterns! LCA provides the ability to easily identify what the most common response patterns were by classifying examinees based on similar response patterns. Using a quantitative method to synthesize response patterns provides a quick and meaningful way of understanding how the majority of examinees are responding to the item, and whether it is in the manner assumed in the initial design of the item and scoring rubric. Let’s say, for example, that after piloting and scoring an item, I observe that 80% of examinees answered the
item incorrectly and 20% correctly. Did so many examinees answer incorrectly because the item measured higher order mathematical ability and thus was more difficult due to the cognitive complexity needed to solve the item? By relying only on the information provided by the score, this is difficult to ascertain. Although traditionally cognitive interviews are conducted to answer this questions, they are time and resource intensive. Digging deeper into how they responded using LCA, may provide the necessary insight more efficiently.

After conducting the LCA, the classes created based on the most typical item responses were compared and contrasted to the scoring rubrics to ascertain whether they accurately reflect examinee responses. Because we are using LCA not to confirm a hypothesis about an underlying latent trait, but rather to provide a summary of typical response patterns in an item, an exploratory approach was used. During this exploratory approach models were tested sequentially with the most parsimonious model tested first, i.e., starting with a model that includes two latent classes then testing three latent classes, four latent classes, etc. The number of classes of examinee response patterns represent specific response patterns common among a number of examinees. Knowing, from the scoring rubric, which response patterns would be awarded points, I could determine whether the scoring rubric accounted for all the possible combinations of correct answers.

As there are currently no common procedures accepted on how to choose the model of classes that best fit the data (Nylund, Asparougov, Muthen, 2007), I decided to use many different indicators including two information criterion indexes, two likelihood ratio tests, followed by a judgment of the relative interpretability of the classes.

The information criterion indexes used were Akaike’s Information Criterion (AIC; Akaike, 1987) and Bayesian Information Criterion (BIC; Schwarz, 1978). They are commonly
used to compare across several different models with the lowest value denoting the best fitting model. Both models are based on the log likelihood of a fitted model but apply different penalties in terms of the number of model parameters and/or sample size. This makes it possible for the two different information criterion indexes to point towards different class solutions as the best fitting model, and is therefore important to examine together.

The likelihood ratio test used were the Lo, Medell, and Rubin (LMR; 2001) and the parametric bootstrap method (BLRT; Mclachlan & Peel, 2000). Differently from the information criterion indexes, the likelihood ratio tests provides significance testing on the difference between two models. A non-significant p-value indicates that the k-1 class model is a better fitting model than k. For example, I were to observe a non-significant p-value for a 4 class solution that would suggest that a 3 class solution exhibited better fit.

Results

As previously states, two exemplar items were chosen from the data described above in use for study 2 to demonstrate how examining response patterns gives us nuanced information not easily observed in the scores. The sample of participating examinees for these specific items are provided below in Table 4.
Table 4. *Examinee Demographic Characteristics for Exemplar Items.*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Exemplar Item 1</th>
<th></th>
<th>Exemplar Item 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>%</td>
<td>N</td>
<td>%</td>
</tr>
<tr>
<td>Total Number of Examinees</td>
<td>231</td>
<td>100.0%</td>
<td>192</td>
<td>100.0%</td>
</tr>
<tr>
<td>School District A</td>
<td>114</td>
<td>49.4%</td>
<td>91</td>
<td>47.4%</td>
</tr>
<tr>
<td>School District B</td>
<td>117</td>
<td>50.6%</td>
<td>101</td>
<td>52.6%</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>73</td>
<td>62.4%</td>
<td>56</td>
<td>55.4%</td>
</tr>
<tr>
<td>Female</td>
<td>44</td>
<td>37.6%</td>
<td>45</td>
<td>44.6%</td>
</tr>
<tr>
<td>Unknown</td>
<td>114</td>
<td>49.4%</td>
<td>101</td>
<td>52.6%</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>68</td>
<td>29.4%</td>
<td>77</td>
<td>40.3%</td>
</tr>
<tr>
<td>Hispanic</td>
<td>130</td>
<td>56.3%</td>
<td>75</td>
<td>39.3%</td>
</tr>
<tr>
<td>African-American</td>
<td>12</td>
<td>5.2%</td>
<td>22</td>
<td>11.5%</td>
</tr>
<tr>
<td>Asian/Pacific Islander</td>
<td>3</td>
<td>1.3%</td>
<td>3</td>
<td>1.6%</td>
</tr>
<tr>
<td>American Indian/Alaskan Native</td>
<td>10</td>
<td>4.3%</td>
<td>6</td>
<td>3.1%</td>
</tr>
<tr>
<td>Multi-Racial</td>
<td>8</td>
<td>3.5%</td>
<td>8</td>
<td>4.2%</td>
</tr>
<tr>
<td>Unknown</td>
<td>0</td>
<td>0%</td>
<td>1</td>
<td>0.5%</td>
</tr>
<tr>
<td>English Language Learner</td>
<td>34</td>
<td>14.7%</td>
<td>10</td>
<td>5.2%</td>
</tr>
<tr>
<td>Special Education</td>
<td>49</td>
<td>21.2%</td>
<td>26</td>
<td>13.5%</td>
</tr>
<tr>
<td>Free/Reduced Lunch</td>
<td>157</td>
<td>68%</td>
<td>97</td>
<td>50.5%</td>
</tr>
</tbody>
</table>
Exemplar item 1: the field trip item. The first item is what we will refer to as “the field trip” item (Figure 10). It was intended as a fourth grade item which asked examinees to identify the correct combination of vehicles needed to take three classes on a field trip. Five options were given, and examinees were asked to “select all that apply”.

![Image of vehicles with numbers of seats and options for combinations]

Figure 10. The field trip item.

The correct answer included selecting options A, C and D. For purposes of scoring, points were allocated based on the following scoring rubric:

- 0 if no correct answer choices were made out of any possible number submitted
- 0 if all five choices were submitted
- 0 if two correct and two incorrect choices were made out of four submitted
- 1 if three correct answer choices were made out of four submitted
- 1 if two correct answer choices were made out of three submitted
- 1 if two correct answer choices were made out of two submitted
- 2 if three answer choices made were correct out of three submitted

Of the 231 responses, 74% were allocated 0 points, 15% allocated one point, and 11% allocated two points. These results as seen in the previously conducted pilot study portrayed to us that this item was particularly difficult for examinees. But, no other information about examinees’ relative performance on the item could be gleaned from only examining the scores. Thus, we
subjected the response patterns for each examinee to a LCA to determine whether our assumptions about the relative difficulty of the items based on the scores was correct.

Based on the BIC, the best class solution was four classes (Table 5). In addition, the LMR-LRT showed non-significance at the five class solution suggesting a four class solution as the best fit. Upon examination of the relative interpretability of the classes, it was concluded that the four class solution was the best fitting model.

Table 5. Model Fit Indices for Classes 2 through 6.

<table>
<thead>
<tr>
<th># of classes</th>
<th>Log-likelihood</th>
<th># of parameters</th>
<th>AIC</th>
<th>BIC</th>
<th>LMR-LRT p-value</th>
<th>BLRT p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>-637.9</td>
<td>11</td>
<td>1297.8</td>
<td>1335.6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>-620.0</td>
<td>17</td>
<td>1274.1</td>
<td>1332.6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>-597.3</td>
<td>23</td>
<td>1240.6</td>
<td>1319.8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>-583.8</td>
<td>29</td>
<td>1225.5</td>
<td>1325.5</td>
<td>.05</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>-572.3</td>
<td>35</td>
<td>1214.7</td>
<td>1335.2</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

In order to better understand the nature of the four classes, response probabilities for each item were examined (Table 6). Each class can be characterized by the following: class 1 had a high probability of selecting option E (74%), class 2 had a high probability of selecting options A (49%), C (100%) and D (60%), class 3 has a high probability of selecting option A (100%), and class 4 had a high probability of selecting option D (100%).

Table 6. Probability of Selecting each Response Option by Latent Class.

<table>
<thead>
<tr>
<th>Option</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Class 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>9</td>
<td>49</td>
<td>100</td>
<td>17</td>
</tr>
<tr>
<td>B</td>
<td>34</td>
<td>5</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>12</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>6</td>
<td>60</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>E</td>
<td>74</td>
<td>4</td>
<td>0</td>
<td>13</td>
</tr>
</tbody>
</table>

Response patterns by each class were then examined in order to further understand the classifications (Table 7). The majority response patterns for each class yield the following
pattern: respondents in class one tended to select only option \( E \) (55%); respondents in class two tended to select options \( A, C \) and \( D \) (37%) and only option \( C \) (31%); respondents in class three tended to select only option \( A \) (95%); and respondents in class four tended to select only option \( D \) (70%). Based on these response patterns it appears that the respondents tended to select either the correct answer (\( A, C, \) and \( D \)) or only one option (\( A, C, D \) or \( E \)).

Table 7. *Frequency of Response Patterns by Class.*

<table>
<thead>
<tr>
<th>Class</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - E</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>26</td>
<td>55</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>47</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 - A, C &amp; D</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>26</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>11</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>22</td>
<td>31</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>71</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 - A</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>41</td>
<td>95</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>43</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 - D</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>9</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>49</td>
<td>70</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>70</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Based on the original scoring rubric only 11% of respondents received full credit (i.e., two points), 15% received partial credit (i.e., one point) and 74% received no credit. In examining the latent classes based on the most likely response patterns, it is evident why so many respondents were allocated zero points. A large number of respondents selected only one option, giving them zero points regardless of whether that option was one of the three possible correct options (e.g., the correct options were A, C and D and the respondent selected option A, but received zero points). Although examinees may be experienced enough with computers and technology so that the response format did not affect how they answered (see Study 1 above), they have most likely been inundated over the years with multiple-choice single answer tests so prominent in U.S. culture. It is possible that they answered in the manner they were accustomed by choosing only one answer, or the item prompt was confusing.

Based on the new knowledge regarding the most likely responses, the rubric was adapted to now reflect the following:

- 0 if no correct answer choices were made out of any possible number submitted
- 0 if all five choices were submitted
- 0 if two correct and two incorrect choices were made out of four submitted
- 1 if one correct answer choice was made and no other choices were submitted
- 2 if three correct answer choices were made out of four submitted
- 2 if two correct answer choices were made out of three submitted
- 2 if two correct answer choices were made out of two submitted
- 3 if three answer choices made were correct out of three submitted

With the adapted rubric that allowed for partial credit for the selection of one single correct answer and no other options selected, the percentage of respondents earning zero points dramatically decreased (see Table 8). Only 26% of respondents earned zero points as compared to 74% under the original rubric. The original scoring rubric made it appear that examinees performance on this item was abysmal, but in reality nearly half (48%) of the respondents were able to identify at least one correct option.
Table 8. *Comparison of Latent Class Membership and Points Earned in Original and Adapted Rubrics.*

| Latent Class | Points Earned | Original Rubric | | Adapted Rubric | |
|--------------|--------------|----------------|----------------|----------------|
|              | 0 | 1 | 2 | 0 | 1 | 2 | 3 | 0 | 1 | 2 | 3 | 0 | 1 | 2 | 3 |
| E (n = 47)   | 47 | 0 | 0 | 47 | 0 | 0 | 0 | 47 | 0 | 0 | 0 | 47 | 0 | 0 | 0 | 47 | 0 | 0 | 0 |
| A, C & D (n = 71) | 23 | 22 | 26 | 1 | 22 | 22 | 26 | 1 | 22 | 22 | 26 | 1 | 22 | 22 | 26 | 1 | 22 | 22 | 26 |
| A (n = 43)   | 43 | 0 | 0 | 2 | 41 | 0 | 0 | 2 | 41 | 0 | 0 | 2 | 41 | 0 | 0 | 2 | 41 | 0 | 0 | 2 |
| D (n = 70)   | 58 | 12 | 0 | 9 | 49 | 12 | 0 | 9 | 49 | 12 | 0 | 9 | 49 | 12 | 0 | 9 | 49 | 12 | 0 |
| TOTAL        | 171 | 34 | 26 | 59 | 112 | 34 | 26 | 59 | 112 | 34 | 26 | 59 | 112 | 34 | 26 | 59 | 112 | 34 | 26 |

**Exemplar item 2: the growth medium item.** A very similar item which showed very different results is the Growth Medium item. It is an Algebra item that asked respondents which statements were true about a specific formula (Figure 11). Eight options were given and respondents were asked to “select all that apply”. The correct answer included selecting options C, E and F. The item was scored dichotomously with one point being given only if the respondents correctly selected the three correct options.

Let \( w_n \) represent the number of cells in the growth medium in week \( n \). Which of these statements are true about the explicit formula for \( w_n \)?

Select all that apply.

- A. \( w_n = 15 + 15 \cdot 2^n - 1 \)
- B. \( w_n = 15 \cdot 15 \cdot 2^n \cdot n \)
- C. \( w_n = 15 \cdot 2^n - 1 \)
- D. \( w_n = \frac{1}{2} \cdot 15 \cdot 2^n - 1 \)
- E. \( w_n = \frac{1}{2} \cdot 15 \cdot 2^n \)
- F. \( n \geq 1 \), where \( n \) is an integer
- G. \( n \geq 1 \), where \( n \) is a real number
- H. \( n \) can be any real number

*Figure 11.* The growth medium item.
As can be seen in Table 9, based on the BIC the best class solution was two classes. In addition, the LMR-LMRT and the BLRT showed non-significance at the three class solution also suggesting a two class solution as the best fit. But, the LMR-LRT was close to non-significance at the two class solution suggesting these data may not fit a multiple class solution. After examining the relative interpretability of the classes, it was determined that a one class solution was the best fitting model. This is most likely due to the high variability in response patterns. There were 92 different response patterns. This suggests that examinees had a very difficult time responding to this item.

<table>
<thead>
<tr>
<th># of classes</th>
<th>Log-likelihood</th>
<th># of parameters</th>
<th>AIC</th>
<th>BIC</th>
<th>LMR-LRT p-value</th>
<th>BLRT p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>-926.3</td>
<td>17</td>
<td>1886.6</td>
<td>1942.1</td>
<td>.03</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>-914.9</td>
<td>26</td>
<td>1881.8</td>
<td>1966.5</td>
<td>.27</td>
<td>.04</td>
</tr>
</tbody>
</table>

**Summary of Findings**

Results from study two suggested that LCA is a useful tool for diagnosing potential issues in the design of items and the design of their scoring rubrics. LCA revealed that the original scoring rubric for exemplar item one did not accurately specify partial credit answers. It identified that many examinees selected only one correct option, but were not given any credit. Not only does this point to poor specification of the scoring rubric, but also a possible design flaw. Examinees may have chosen one and only one answer because they did not understand the directions to “select all that apply”.

LCA revealed for exemplar item 2 a myriad of response patterns. Examinees may have been unable to identify a correct response because the item design. For example, the number of response options for the student to select was 8. This means that there were 40,320 possible response patterns. Furthermore, because the scoring rubric only permitted for the identification
of one correct response, no partial credit, that gives examinees 0.00000025% chance of selecting the correct response.

**Chapter 7. Discussion**

Study 1 provides evidence that even if some item formats are more difficult to use, they do not significantly prevent an examinee’s ability to answer an item correctly. Therefore, it is likely that examinees are accustomed to using these item formats and technologies included in them (selecting, drag-and-drop, fill-in-the-blank). However, this empirical evidence does not apply to all TEIs. The computer-based items used in this study were closer to technology-enabled than technology-enhanced. More akin to technology-enhanced are the items presented in the introduction of this paper (Figures 1, 2, and 3). They bring forth appreciably more complex integrations of technology that have yet to be methodically evaluated.

For this reason, cognitive load theory ought to be included as a central tenant to assessment design frameworks. Having a cognitive model that explains, at least theoretically, how examinees interact with complex TEIs is essential; just as having a cognitive model underlying what the scores are intent to portray. Evidence Centered Design framework is intended to provide the language, concepts, and knowledge representations for designing and delivering educational assessments. Cognitive load theory is a conceptual tool that can be used by developers when building a task model—a description of the environment in which students say, do, or make something to provide evidence. Having a sound theoretical model to make more explicit design parameters in computer-based testing may help streamline authoring, implementation, and evaluation (Mislevy & Haertel, 2006).

If we want to maintain the idea that TEIs make enhancements available that better measure more complex skills, test developers may have to provide more explicit evidence to
support that the TEIs do just that. And that the TEIs do not, on the other hand, produce
extenuating circumstances of cognitive load inhibiting examinees from being able to demonstrate
what they know and are able to do. That is, to better measure complex skills, we must not also
inadvertently introduce construct irrelevant variance. Regardless of whether the content of a
TEI reflects the relevant knowledge, skills, and abilities to be measured, it is conceivable that the
depletion of cognitive resources due to poor item design may introduce construct irrelevant
variance. Examinees may make errors that are unrelated to lack of knowledge as the observed
scores would imply, but rather to unrecognized variations in cognitive load (Ayres, 2001).

Study 2 also has implications for assessment design as it demonstrates the utility of
latent class analysis as an evaluative tool that can be effectively used to flag flaws in the design
of the items and scoring. As using cognitive load theory may provide a conceptual framework
for the task model in evidence centered design, latent class analysis is a methodological tool that
can be applied to the evidence model—evaluation and measurement bridging the student and
task models. Latent class analyses, by identifying and evaluating salient aspects of an
examinee’s performance, may provide an efficient way of updating our beliefs about an
examinee after observing what they do.

With the massive number of examinees typically tested in many large-scale assessments,
it is nearly impossible to poorly performing test items without a toolbox of statistical tools (e.g.,
differential item functioning methods, classical test theory and item response theory models) to
analyze and synthesize examinee responses. Typically, poorly performing items are flagged with
item response theory models (citation here). But, these models offer little insight into what the
potential design are and how they can be mitigated. Latent class analysis, on the other hand,
appears to be a useful tool for raising awareness of the design issues that may be contributing to the various responses patterns in the empirical data derived from field trials of TEIs.

The investigations reported above on the exemplar items demonstrated just that. In the absence of latent class analysis, we would not have realized that there were significant problems with the scoring rubric. Additionally, latent class analysis helped highlight the need to examine further the wording of the prompt given it appeared that examinees thought only to select one answer instead of “all that apply”. Although a similar item format, latent class analysis of exemplar item 2 indicated that examinees were likely completely confused. We may not be able to tell exactly why, but we now know to revise the item being more mindful of where the confusion may have presented itself. These pitfalls can then be attended to during the design of new items to improve proficiency in the process of design.

Chapter 8. Conclusion

Evidence provided from both studies illuminates the immediate need to further research computer based items that are beginning to be used widely in high stakes, large-scale assessments. In an effort to move away from traditional multiple choice items and toward more authentic measurement by incorporating technology based item features, we may be affecting how examinees respond to the item due to inadvertent increases in cognitive load. With the implementation of new state accountability tests aligned to the common core standards such as PARCC and SBAC, buzz can be heard throughout the country about the difficulty of the tests and scores are reflecting the higher bar set for proficiency. But, it is important to ask whether difficulty of the TEIs is being inflated by cognitive load demands.

Furthermore, careful and considered application of cognitive load theory to item design may also unlock the secrets of how to make test items (TEIs in particular) more or less difficult.
This design challenge is hampering the design of many large-scale assessment programs, such as NAEP, PISA, PARCC and SBAC, resulting in poor measurement at the lower ends of the ability range—by reducing the cognitive load, through design, we may be able to create less difficult items without watering down the content.

**Limitations**

A limitation of these studies includes the lack of variables relating the item characteristics to better tease out what characteristics introduce construct irrelevant variance. In study one, the multi-level model suggested that about 50% of the variance in scores was between-items. Even after adding in student characteristics, this variance attributed to item characteristics was still present. Therefore, there is unexplained variation between items that the item response format (e.g., multiple-choice-single-answer and drag-and-drop) was not picking up.

There is also a limitation with the data which constrained what type of modeling I could use. Matrix sampling was not implemented that would allow for examining the full sample of examinees. Instead groups of items at each grade level were given to different groups of examinees. This made it difficult to do any IRT modeling.

Lastly, the fact that these were prototype items is a limitation to the study. Therefore, these items could be argued as being inherently flawed. These items were first blush designs being piloted to examinees to understand whether they were “good” items to include in an operational test. Although this is true, it is important that we should carefully examine the prototype items and their flaws before including them in an operational test.

**Future Research**

I recommend that similar investigations to the current study as well as more thorough investigations with experimental manipulation, be consistently included in the process of item
design. Future research involving experimental manipulation is necessary for understanding how item characteristics impact how examinees respond to an item. I suggest obtaining a set of items from multiple grade levels that will be or are currently being used in large-scale assessments. The independent features can be manipulated individually to assess the impact in variations. For example, an item can be manipulated solely in its response interface to determine whether, all other things being equal, selecting or drag-and-drop interfaces influences performance. Further investigations can help to create relevant guidelines for item writers; similar to those developed out of research on multiple choice items.
Appendix A. Task Models for Design of Mathematics Prototype Items

Type I Task Models: Tasks that generate evidence for sub claims A, B and/or E, and that do not generate evidence for sub claims C or D. They are machine scoreable including innovative, computer-based formats. Balance conceptual understanding, procedural knowledge and brief applications. And can be practice-forward.

Type II Task Models: Tasks that generate evidence for claim C. Tasks are hand scored, or machine scored with innovative computer-based formats, or a combination. Each task calls for written arguments/justifications, critique of reasoning, or precision in mathematical statements. Tasks may be practice-forward in other ways as well.

Type II Task Models: Tasks that generate evidence for claim D. Tasks are hand scored, or machine scored with innovative computer-based formats, or a combination. Each task involves a given real-world context or scenario. Tasks are practice-forward, highlighting modeling and potentially involving some or all of mathematical practices 1, 2 5, 7 & 8.

Master Claim: On-Track for College and Career Readiness. The student is on-track/not on track for college and career readiness in mathematics. The student solves grade-level/course-level problems in mathematics as a set forth in the standards for Mathematical Content with connections to the Standards for Mathematical Practice.

Sub-claim A: Major Content with Connections to Practices. The student solves problems involving the Major Content for her grade/course with connections to the Standards for Mathematical Practice.

Sub-claim B: Additional and Supporting Content with Connections to Practices. The student solves problems involving grade/course with connections to the Standards for Mathematical Practice.

Sub-claim C: Highlighted Practices MP.3,6 with Connections to Content: expressing mathematical reasoning. The student expresses grade/course-level appropriate mathematical reasoning by constructing viable arguments, critiquing the reasoning of others, and/or attending to precision when making mathematical statements.

Sub-claim D: Highlighted Practice MP.4 with Connections to Content: modeling/application. The student solves real-world problems with a degree of difficulty appropriate to the grade/course by applying knowledge and skills articulated in the standards for the current grade/course (or, for the more complex problems, knowledge and skills articulated in the standards for previous grades/courses), engaging particularly in the Modeling practice, and where helpful making sense of problems and persevering to solve them (MP.1), reasoning abstractly and quantitatively (MP.2), using appropriate tools strategically (MP.5), looking for and making use of structure (MP.7), and/or looking for and expressing regularity in repeated reasoning (MP.8).
Sub-claim E: Fluency in applicable grades (3-6): The student demonstrates fluency as set forth in the Standards for Mathematical Content in her grade.

Mathematical Practices:
1. Make sense of problems and persevere in solving them
2. Reason abstractly and quantitatively
3. Construct viable arguments and critique the reasoning of others
4. Model with mathematics
5. Use appropriate tools strategically
6. Attend to precision
7. Look for and make use of structure
8. Look for and express regularity in repeated reasoning
## Appendix B. Mapping of Prototype Tasks onto CCSS and Mathematical Practices

### Mapping of Prototype Tasks onto CCSS

<table>
<thead>
<tr>
<th>CCSS Standards</th>
<th>Tasks (N)</th>
<th>Discrete Items (N)</th>
<th>Samples (N)</th>
<th>Examinees per sample (N)¹</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.OA.3</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>184</td>
</tr>
<tr>
<td>3.OA.5</td>
<td>3</td>
<td>6</td>
<td>2</td>
<td>184, 170</td>
</tr>
<tr>
<td>3.MD.7a</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>184</td>
</tr>
<tr>
<td>3.MD.7b</td>
<td>5</td>
<td>9</td>
<td>3</td>
<td>184, 232, 232</td>
</tr>
<tr>
<td>3.MD.7c</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>184</td>
</tr>
<tr>
<td>3.MD.7d</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>184</td>
</tr>
<tr>
<td>3.NF.2a</td>
<td>6</td>
<td>21</td>
<td>2</td>
<td>184, 169</td>
</tr>
<tr>
<td>3.NF.2b</td>
<td>6</td>
<td>21</td>
<td>2</td>
<td>184, 169</td>
</tr>
<tr>
<td>3.NF.3a</td>
<td>6</td>
<td>21</td>
<td>2</td>
<td>184, 169</td>
</tr>
<tr>
<td>3.NF.3b</td>
<td>6</td>
<td>21</td>
<td>2</td>
<td>184, 169</td>
</tr>
<tr>
<td>3.NF.3c</td>
<td>3</td>
<td>14</td>
<td>2</td>
<td>184, 169</td>
</tr>
<tr>
<td>3.G.2</td>
<td>3</td>
<td>14</td>
<td>2</td>
<td>184, 169</td>
</tr>
<tr>
<td>3.MD.7b</td>
<td>5</td>
<td>9</td>
<td>3</td>
<td>184, 232, 232</td>
</tr>
<tr>
<td>3.MD.8</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>232, 232</td>
</tr>
<tr>
<td>4.NBT.4</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>233, 225</td>
</tr>
<tr>
<td>4.NBT.5</td>
<td>6</td>
<td>8</td>
<td>2</td>
<td>233, 225</td>
</tr>
<tr>
<td>4.OA.3</td>
<td>6</td>
<td>8</td>
<td>2</td>
<td>233, 225</td>
</tr>
<tr>
<td>4.NBT.1</td>
<td>4</td>
<td>9</td>
<td>2</td>
<td>234, 229</td>
</tr>
<tr>
<td>4.NBT.2</td>
<td>4</td>
<td>9</td>
<td>2</td>
<td>234, 229</td>
</tr>
<tr>
<td>4.NBT.3</td>
<td>4</td>
<td>9</td>
<td>2</td>
<td>234, 229</td>
</tr>
<tr>
<td>6.RP.3a</td>
<td>3</td>
<td>10</td>
<td>1</td>
<td>221</td>
</tr>
<tr>
<td>6.EE.9</td>
<td>9</td>
<td>19</td>
<td>2</td>
<td>221, 215</td>
</tr>
<tr>
<td>6.NS.6c</td>
<td>3</td>
<td>11</td>
<td>2</td>
<td>214, 189</td>
</tr>
<tr>
<td>6.NS.7b</td>
<td>3</td>
<td>11</td>
<td>2</td>
<td>214, 189</td>
</tr>
<tr>
<td>6.NS.7c</td>
<td>3</td>
<td>11</td>
<td>2</td>
<td>214, 189</td>
</tr>
<tr>
<td>6.RP.1</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>198, 215</td>
</tr>
<tr>
<td>6.EE.3</td>
<td>3</td>
<td>6</td>
<td>1</td>
<td>208</td>
</tr>
<tr>
<td>7.NS.3</td>
<td>3</td>
<td>6</td>
<td>2</td>
<td>184, 178</td>
</tr>
<tr>
<td>7.RP.1</td>
<td>2</td>
<td>6</td>
<td>2</td>
<td>181, 182</td>
</tr>
<tr>
<td>7.EE.3</td>
<td>3</td>
<td>8</td>
<td>2</td>
<td>181, 182</td>
</tr>
<tr>
<td>7.EE.1</td>
<td>6</td>
<td>14</td>
<td>2</td>
<td>179, 168</td>
</tr>
<tr>
<td>7.EE.4a</td>
<td>3</td>
<td>8</td>
<td>2</td>
<td>166, 168</td>
</tr>
<tr>
<td>7.NS.3</td>
<td>3</td>
<td>6</td>
<td>2</td>
<td>184, 178</td>
</tr>
<tr>
<td>7.RP.3</td>
<td>3</td>
<td>6</td>
<td>2</td>
<td>179, 161</td>
</tr>
<tr>
<td>G.CO.5</td>
<td>3</td>
<td>7</td>
<td>2</td>
<td>226, 205</td>
</tr>
<tr>
<td>F.BF.3</td>
<td>3</td>
<td>6</td>
<td>2</td>
<td>226, 205</td>
</tr>
<tr>
<td>8.G.3</td>
<td>3</td>
<td>6</td>
<td>2</td>
<td>226, 205</td>
</tr>
</tbody>
</table>
Note: Some tasks have missing data. Therefore, the number of examinees was calculated by taking the highest number of examinees within a sample who completed a task. For example, if 184 examinees completed task 3A1 and 177 completed task 3A6, then the number of examinees completing the tasks would be 184.

“Task 3AB3 is a special case as two third grade samples took this task totaling 348 examinees.

<table>
<thead>
<tr>
<th>Mathematical Practices</th>
<th>Tasks (N)</th>
<th>Discrete Items (N)</th>
<th>Grade Levels Tested</th>
<th>Examinees per Sample (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MP.1</td>
<td>6</td>
<td>10</td>
<td>4 &amp; 7</td>
<td>233, 225, 184, 178</td>
</tr>
<tr>
<td>MP.2</td>
<td>18</td>
<td>37</td>
<td>3, 4, 6 &amp; Alg</td>
<td>184, 170, 233, 225, 198, 215, 221, 218</td>
</tr>
<tr>
<td>MP.3</td>
<td>13</td>
<td>35</td>
<td>3, 4, 7 &amp; Alg</td>
<td>184, 168, 234, 229, 179, 161, 226, 205</td>
</tr>
<tr>
<td>MP.4</td>
<td>19</td>
<td>54</td>
<td>3, 4, 6, 7 &amp; Alg</td>
<td>184, 232, 223, 221, 166, 168, 221, 218</td>
</tr>
<tr>
<td>MP.5</td>
<td>19</td>
<td>54</td>
<td>3, 4, 6, 7 &amp; Alg</td>
<td>184, 232, 223, 221, 166, 168, 221, 218</td>
</tr>
<tr>
<td>MP.6</td>
<td>13</td>
<td>35</td>
<td>3, 4, 6 &amp; Alg</td>
<td>184, 169, 234, 229, 208, 226, 205</td>
</tr>
<tr>
<td>MP.7</td>
<td>14</td>
<td>37</td>
<td>3, 4 &amp; Alg</td>
<td>184, 170, 232, 222, 226, 205</td>
</tr>
<tr>
<td>MP.8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Appendix C. Study 1 Statistical Model

Level-1 Model: The Student Level Model

Within each item, student item scores as a function of student-level predictors plus a random student level error was modeled. As can be seen below, the level-1 model indicates that the item scores were predicted by examinees’ gender (Gender), race/ethnicity (Race), Free/reduced lunch status (FRLunch), and prior math scores (ZMath) and prior reading scores (ZRead).

\[
Score_{ij} = \beta_{0j} + \beta_{1j}Gender_{ij} + \beta_{1j}Race_{ij} + \beta_{1j}FRLunch_{ij} + \beta_{1j}ZMath_{ij} + \beta_{1j}ZRead_{ij} + e_{ijk}
\]

In the above models, you will notice the subscripts \( i \) and \( j \). The subscript \( i \) indexes examinees within items where \( j \) indexes items. The notation \( \beta \) represents the coefficients in this level-1 regression model. Where \( \beta_{0jk} \) is the intercept for item \( j \) and \( \beta_{pj} \) are the corresponding level-1 coefficients that indicate the direction and strength of association between each student characteristics and the outcome in item \( jk \). Lastly, \( e_{ij} \) is a level-1 random effect that represents the deviation of student \( ij \)’s score from the predicted score based on the student-level model.

Level-2 Model: The Item Level Model

Each of the regression coefficients in the student-level model (including the intercept) can be viewed as either fixed, nonrandomly varying, or random. These possibilities lead to the below formulation of the model for variation among items. The level-2 model indicates that each item effect can be predicted by item type (Type).

\[
\beta_{0j} = \gamma_{p0} + \gamma_{p1}Type_{ej} + r_{pj}
\]

\[
\beta_{1j} = \gamma_{p0} + \beta_{p1}Type_{jk} + r_{pj}
\]

For the level-2 model, \( \gamma \) represents the regression coefficients. Where \( \gamma_{p0} \) is the intercept in modeling the item effect \( \beta_{pj} \). And \( \gamma_{pq} \) is the corresponding coefficient that represents the direction and strengths of association between item characteristic and \( \gamma_{pj} \). Lastly, \( r_{pj} \) is a level-2 random effect that represents the deviation of item \( j \)’s level-1 coefficient, \( \beta_{pj} \), from its predicted value based on the item-level model.
Appendix D. Study 1 Syntax for Multi-Level Model

STATA/SE 14.0 (StataCorp, 2015)

The intercept-only model:

melogit Score | | Item

The means-as-outcomes model:

melogit Score | | Item: dnd fib mcma

Random-regression coefficients model:

melogit Score gender ethnicity free_red_lunch ZMath ZRead | | Item:

Intercepts- and slopes-as-outcomes model:

melogit Score gender ethnicity free_red_lunch ZMath ZRead | | Item: dnd fib mcma
Appendix E. Study 2 Syntax for Latent Class Analyses

MPLUS 6 (Muthen & Muthen, 2007)

Title:
LCA for item 4A2a

Data:
File is "E:\flash drive 20131105\Dissertation\Data\Master Files\Pilot Data_Dissertation_MASTER_20150528.csv";

Variable:
NAMES ARE id A_4A2a B_4A2a C_4A2a D_4A2a E_4A2a;
IDVARIABLE = id;
USEVAR = A_4A2a B_4A2a C_4A2a D_4A2a E_4A2a;
CATEGORICAL = A_4A2a B_4A2a C_4A2a D_4A2a E_4A2a;
MISSING ARE ALL (999);
CLASSES = c(4);

ANALYSIS:
type = mixture;

MODEL:
%Overall%

SAVEDATA:
file is Dissertation4A2a _4c.txt;
save is cprob;
format is free;

OUTPUT:
tech10 tech11 tech14;

PLOT:
type = plot3;
series is A_4A2a (1) B_4A2a (2) C_4A2a (3) D_4A2a (4) E_4A2a (5);
References


StataCorp. 2015. *Stata Statistical Software: Release 14*. College Station, TX: StataCorp LP.


Rudner, L. (2010). *Implementing the Graduate Management Admission Test computerized adaptive test*. In W. van der Linden & C. Glas (Eds.), *Elements of adaptive testing* (pp. 151-165), New York, NY: Springer.


