Comparing cognitive models of domain mastery and task performance in algebra: validity evidence for a state assessment

Zachary Borden Warner

University at Albany, State University of New York, zacharybwarner@yahoo.com

The University at Albany community has made this article openly available. Please share how this access benefits you.

Follow this and additional works at: https://scholarsarchive.library.albany.edu/legacy-etd

Part of the Educational Assessment, Evaluation, and Research Commons

Recommended Citation

https://scholarsarchive.library.albany.edu/legacy-etd/1039

This Dissertation is brought to you for free and open access by the The Graduate School at Scholars Archive. It has been accepted for inclusion in Legacy Theses & Dissertations (2009 - 2024) by an authorized administrator of Scholars Archive.
Please see Terms of Use. For more information, please contact scholarsarchive@albany.edu.
COMPARING COGNITIVE MODELS OF DOMAIN MASTERY
AND TASK PERFORMANCE IN ALGEBRA:
VALIDITY EVIDENCE FOR A STATE ASSESSMENT

by

Zachary B. Warner

A Dissertation
Submitted to the University at Albany, State University of New York
in Partial Fulfillment of
the Requirements for the Degree of
Doctor of Philosophy

School of Education
Department of Educational and Counseling Psychology
2013
Abstract

This study compared an expert-based cognitive model of domain mastery with student-based cognitive models of task performance for Integrated Algebra. Interpretations of student test results are limited by experts’ hypotheses of how students interact with the items. In reality, the cognitive processes that students use to solve each item may be very different than those assumed by the test developers. Without knowledge of how students are actually operating on these test items and the cognitive processes they use, interpretations from the results of the examination may be compromised. In this study, I investigated the cognitive processes that students used when solving nine multiple-choice integrated algebra items from a large-scale examination. Once elicited and recorded through think-aloud protocols, the students’ cognitive processes were synthesized into item-level cognitive models of task performance and these models were compared with the less specific expert-based cognitive model of domain mastery currently used to develop the examination. As the level of agreement between these two was very high, this comparison provided evidence for the validation of inferences from test scores. The utility of this information for validation and the implications for test development and score reporting are discussed.
ACKNOWLEDGEMENTS

I certainly could not have completed this dissertation, or my doctoral program as a whole, without the support of my family.

Special thanks to Victoria, Mary Beth, Holly, and Angela for their friendship and guidance these past few years.

Candy Shyer has been amazing in giving me opportunities to grow as a professional and has always offered help along the way.

Heidi, Bob, and Kristen: their willingness to devote the time to this work and take an active interest in my future will never be forgotten.

My wife Michele has endured everything that goes along with being married to a graduate student and shown nothing but love though it all.

For I am convinced… (Romans 8:38-39)
# TABLE OF CONTENTS

## CHAPTER 1: INTRODUCTION AND LITERATURE REVIEW ................................................................. 1

Research Questions .................................................................................................................... 2
Review of the Literature ............................................................................................................. 3
Cognitive Models in Educational Assessment .............................................................................. 3
   Categories of Cognitive Models .......................................................................................... 6
   Relationships between Types of Cognitive Models .............................................................. 12
Cognitive Models as Validity Evidence ..................................................................................... 13
Developing Cognitive Models .................................................................................................. 18
   Retrofitting Cognitive Models .......................................................................................... 23
Cognitive Models in Mathematics ............................................................................................ 24
   Cognitive Models of Task Performance in Mathematics ..................................................... 26

## CHAPTER 2: METHOD .................................................................................................................. 31

Participants ............................................................................................................................... 31
Instrumentation ......................................................................................................................... 34
   June 2009 Regents Examination in Integrated Algebra ....................................................... 34
   Items selected for study ...................................................................................................... 37
   Integrated Algebra Test Specifications ............................................................................. 39
Data Collection .......................................................................................................................... 41
Data Analysis .............................................................................................................................. 43
   Construction of cognitive models of task performance ..................................................... 43
   Comparison of cognitive models of domain mastery and task performance .................... 47

## CHAPTER 3: RESULTS ............................................................................................................... 48

Research Questions 1 and 2 ...................................................................................................... 49
Research Question 3 .................................................................................................................. 65
Research Question 4 .................................................................................................................. 71

## CHAPTER 4: DISCUSSION ......................................................................................................... 75

Implications for Test Development ........................................................................................... 75
   Implications for item writing ............................................................................................. 78
   Implications for Validation ................................................................................................. 80
   Implications for Score Reporting ....................................................................................... 83
Limitations ................................................................................................................................. 85
Future Research Directions ..................................................................................................... 88

## REFERENCES .......................................................................................................................... 91

## APPENDIX A: INTERVIEW PROTOCOL ...................................................................................... 108

## APPENDIX B: CONSENT DOCUMENTATION .......................................................................... 112

## APPENDIX C: STUDENT ASSENT FORM .................................................................................. 114

## APPENDIX D: SCRIPT FOR RECRUITMENT OF STUDENT PARTICIPANTS ......................... 116

## APPENDIX E: INTEGRATED ALGEBRA ITEMS SELECTED FOR STUDY ............................... 117

## APPENDIX F: ITEM ALIGNMENT MAP OBTAINED FROM NYSED ........................................ 120

## APPENDIX G: CODING STRATEGY, POLICY, AND PROCESS FOR CALIBRATION ............. 122
Chapter 1: Introduction and Literature Review

Interpretations of student results on the New York State Regents Examination in Integrated Algebra are constrained by experts’ hypotheses of how students interact with the test items. The items are designed to measure performance indicators which are explanations of the skills that experts believe students need to correctly solve a particular item (or problem). In reality, the cognitive processes that students use to approach each item might be very different than those assumed by the test developers. Elimination or other test-wise strategies or even simple guessing all challenge the assumption that success on any given item can be interpreted as mastery over the aligned content. Without knowledge of how students are actually operating on these test items and the cognitive processes they use, evidence for the validity of interpretations of the examination might be undermined. The National Research Council (2001) stipulated that all tests, especially those with high stakes attached, be based on empirically developed and validated models of student cognition. Given that this examination is used as a high school commencement exam (i.e., to satisfy diploma requirements), it seems important that interpretations are as evidence-based as possible.

The purpose of this study was to gather validity evidence for inferences about student achievement and conceptual understanding drawn from the results of the Regents Examination in Integrated Algebra. This information can also be used to inform future test development efforts. Although a great deal of evidence is gathered by New York State during validation, an analysis of the cognitive processes expected and elicited is not a part of these efforts. In this study, think-aloud protocols were used to develop cognitive models of students’ task performance for test items. These cognitive models display the
specific skills that students employ when solving each multiple-choice item. Each task performance model was compared with the performance indicator the item was written to measure. The collection of performance indicators was assumed to represent an expert-generated cognitive model of domain mastery, within which a test specification cognitive model is nested (Leighton & Gierl, 2007a). The comparison of the two models represents an evaluation of how often expert-expected cognitive processes match the processes used by students when solving the problem. Alignment between these two models suggests that the current interpretations made from the results of the Integrated Algebra exam are appropriate and would bolster the argument for the construct and content validity of these interpretations. However, if the actual cognitive processes deviate from the expected, interpretations made from the results of the exam are called into question because the item is not actually measuring what it is purporting to measure. If test scores reflect what the expert is expecting them to do rather than what the students are actually doing, the interpretations that can be made from the results of the test may be limited.

**Research Questions**

In this study, I sought to answer four research questions related to students’ cognitive processes that are elicited by the Regents Examination in Integrated Algebra:

1. What are the cognitive processes students use to solve nine items selected from the Integrated Algebra Regents examination?
2. How can students’ cognitive processes be represented in cognitive models of task performance?
3. To what extent do the cognitive processes used by students (i.e., the task performance cognitive models) align with the performance indicators expected by content experts for each item (i.e., the domain mastery cognitive model)?

4. To what degree does the construct validity evidence provided by examination of expert-expected and actual student processes support the claims of the exam?

**Review of the Literature**

The purpose of the following chapter is to provide a review of key literature associated with the present study. In this review, I introduce the relevant background information that informs the context of the study. Furthermore, I establish the need for the current line of inquiry in light of previous investigations into cognitive modeling of mathematics for assessment development purposes. To begin the first section, I operationally define the term *cognitive model* and discuss the utility of cognitive models in educational assessment. Next, I introduce processes for the development of cognitive models to provide background for the methodology of the current study. A review of studies that use cognitive models in validation efforts for assessment follows, with an eye towards illustrating the utility of cognitive model-based validity evidence in these large-scale testing programs. Finally, I introduce cognitive models of student and expert processes in order to demonstrate the extent of current research efforts for each type.

**Cognitive Models in Educational Assessment**

The term “cognitive model” emerged from the field of computer science where it referred, and still refers, to some representation of human problem solving, often mathematical or computational. Since this initial appearance, Busemeyer and Diederich (2010) reported that cognitive models have been applied to studies in psychology,
neuroscience, economics, and other fields beyond their origin. The same authors noted that it is important to differentiate the concept of models of cognition from conceptual, statistical, or neural models that may seem to be similar. They have suggested that the distinction lies in the purpose of the cognitive model as interpreted within the field of cognitive science. This discipline focuses on understanding the processes undertaken by the brain when engaging in complex tasks. Thus, a cognitive model is concerned with explaining these processes or the interaction among multiple processes.

In their seminal work on the foundations of educational assessment, the National Research Council (NRC, 2001) asserted that every assessment requires a “model of how students represent knowledge and develop competence in the subject domain [being assessed]” (p. 2). This group went on to require that such a cognitive model would be based on empirical studies of learning within the specific content domain, would distinguish between beginning and expert learners, would express the variety of ways that students may develop knowledge in the domain, and would be flexible in order to allow assessment designers to select various subsets of the model to be used for different assessment purposes. Leighton and Gierl (2007a) argued that the NRC assumption that all assessments are based on a cognitive model of learning is often not realized in practice. Particularly, many large-scale assessments do not use empirical studies of learning as a basis for test development. Thus, if the NRC’s proposal that every assessment be based on a cognitive model was to be realized, the defining characteristics of a cognitive model had to be relaxed in order to include these large-scale tests. According to Leighton and Gierl (2007a), the term cognitive model describes a “simplified description of human problem solving on standardized educational tasks,
which help to characterize the knowledge and skills students at different levels of learning have acquired to facilitate the explanation and prediction of students’ performance” (p. 6). They have qualified this definition by limiting it to the field of educational measurement in order to “broadly…characterize the knowledge and skills examinees require to solve items on assessments (Leighton, & Gierl, 2011, p. 49).

Mislevy (1994, 1996) referred to the process of collecting evidence to support inferences about student abilities as reasoning from evidence. The NRC (2001) noted that this process has three distinct components: cognition, observation, and interpretation. To define each of these components, the NRC emphasized two additional elements beyond their call for assessments to be based on cognitive models. These elements are: 1) tasks that allow students to demonstrate their knowledge, and 2) a method for drawing inferences from these performances. These elements were adapted into the assessment triangle (Figure 1). Leighton (2004) defined each vertex as follows:

The cognition vertex indicates that every test form reflects, either explicitly or implicitly, a cognitive mode of knowledge and skills required for successful performance in the domain of the test. The observation vertex indicates that every test item operationalizes this cognitive model by functioning as a stimulus to elicit the relevant knowledge and skills in students. The interpretation vertex indicates that every test score is based on a psychometric method that is used to summarize test performance and generate inferences from that performance to a larger domain of behavior (p.7).
The triangle design was chosen to emphasize the importance of congruence and alignment of vertices. Without each of these three attributes, any inferences made from a student’s test performance may not stand up to scrutiny. For example, a weak cognitive model can lead to difficulty in selecting appropriate performance tasks or issues in accurately reporting student achievement.

**Categories of Cognitive Models.** Leighton (2004) also defined three distinct categories of cognitive model and commented on the inferences that can be made from each type. The categories of models, from simplest to most complex, are test specification, domain mastery, and task performance. Leighton and Gierl (2007a) expanded on each category of model and classified each hierarchically in terms of specificity of knowledge and skills and insight into students’ thought processes.

The cognitive model of *test specifications* is also sometimes called a test blueprint and represents “an explicit plan that guides test construction” (Thorndike & Thorndike-Christ, 2010, p. 156). Commonly, this model is generated using a two-way matrix of content and skills to be assessed and then used to generate items that are representative of...
the specified domain. The majority of large-scale assessments are developed based on test specifications that list content and (sometimes) skills required for success in the domain (Downing, 2006; Ferrara & DeMauro, 2006). Thorndike and Thorndike-Christ (2010) stressed the importance of linking the content with the cognitive processes that the items are meant to elicit. In certain state-level, standards-based assessments (e.g., New York), cognitive processes have been represented by performance indicators for the content standards being tested. Additionally, weighting allows emphasis to be place on certain content or processes by assigning more points to content or processes that are more highly valued within the domain.

Leighton and Gierl (2007a) applaud the convenience and simplicity of the cognitive model of test specifications but caution that the inferences drawn from such a model must remain simple as well. The performance indicators represented in the specifications are rarely verified to see if students’ actual mental processes are aligned with the expected processes espoused by content experts during test development (Schmeiser & Welch, 2006). In fact, several studies that examined the processes students use to solve test items have suggested that the alignment with the processes expected by content experts is often poor (Baxter & Glaser, 1998; Gierl, 1997; Hamilton, 1994; Hamilton, Nussbaum, & Snow, 1997). In their study of science assessments, Baxter and Glaser (1998) found that students often interpreted complex content in a superficial manner, missing the content-rich experience that experts had planned for them. Additionally, Lomask, Baron, Greig, and Harrison (1992) discovered a situation where a test item and subsequent scoring procedure allowed students to receive credit for knowledge they did not need to display in their responses. Because the teachers assumed
that students held a certain scientific understanding, the students’ responses were misinterpreted as demonstrating knowledge that they did not possess. Yarroch (1991) found similar results after conducting think-aloud protocols with 20 students and concluded that “[i]t is all too easy to assume that the students we teach know and understand more than they actually do” (p. 628).

One possible explanation for this mismatch is that the people developing the test specifications, which are often established at the policy level, may not be those who are best-suited to ensure the specifications represent the best thinking on student cognition within the content area. According to Leighton and Gierl (2007a), it is unlikely that student thinking processes differ completely from those processes expected by experts, or what the authors later dubbed “cognitive intentions” (2011, p. 51). Still, even slight deviations call the validity of inferences into question due to the introduction of construct-irrelevant variance (Messick, 1989, 1995). That is, student performance may be attributed to qualities separate from the construct intended for assessment. In addition, the knowledge and skills measured by tests developed from specification models are often very general and do not allow for several items targeting the same skill due to the need to cover a great deal of content material.

It is typical for a cognitive model of test specifications to be developed by a group of content and assessment experts brought together for the purpose of designing a test. This development occurs most often in large-scale assessments such as those used for accountability under No Child Left Behind (2001; Ferrara & DeMauro, 2006). Experts’ recommendations are collected by the agency administering the test and a final decision is made on content coverage and, often, item types (e.g., NYSED, 2009b). Thus, models
of test specifications are entirely expert-dependent and are based on the beliefs of experienced people about what the test should look like and, in general terms, what content it should assess. If this group does not include members who have experience within the content area and can hypothesize on models of student cognition, there is a possibility that the resulting specifications will not allow for the measurement of the intended knowledge and skills.

The cognitive model of domain mastery requires identification of the knowledge and skills necessary to signify expertise within a specified content or achievement domain. Svetina, Gorin, and Tatsuoka (2011) referred to these models as “skills-based models” (p. 3). Curriculum-based tests which are designed to comprehensively address the knowledge and skills of a particular content area at a certain grade level typically rely on cognitive models of domain mastery during development (Svetina et al., 2011; Wilson & Sloane, 2000). There is an emphasis on the match between instruction and assessment because the curriculum must be developed concurrently with the assessments to ensure that the tests are representative and comprehensive of what is to be taught in the classroom.

Leighton and Gierl (2007a) classified the domain mastery model as more complex than the cognitive model of test specifications due to the inclusion of detailed information on the knowledge and skills necessary to demonstrate expertise of the specified content. The authors also approved of the multiple opportunities for students to receive feedback based on their progress through the curriculum. Yet, two disadvantages are noted. First, identifying all of the necessary knowledge and skills that define a content domain is time consuming, as is testing students in multiple iterations to appropriately assess student
achievement. The second disadvantage of domain mastery models is that accepting the behavioral outcome of item response as a cognitive process does not correspond with the diagnostic capabilities of cognitive models. That is, a teacher interpreting the results of a test based solely on a cognitive model of domain mastery has no framework to help him or her understand what an incorrect response signifies in terms of underlying cognitive processes.

Cognitive models of domain mastery incorporate an additional dimension beyond those of test specifications. Domain mastery models include the content and item overview that make up the test specifications model but also list the skills necessary to demonstrate mastery of the content assessed. Gorin (2006) likened this cognitive model to the student model described in Evidenced Centered Design (ECD; Mislevy, Steinberg, & Almond, 2003) framework for test development. However, Gorin noted that the ECD student model focuses on inferences about skills rather than listing the skills, as is done in many large-scale tests. Additionally, Shute, Hansen, and Almond (2008) suggested that ECD falls more in line with “the kind of thinking ordinarily done by expert assessment developers” (p. 294).

The final category of cognitive model discussed by Leighton and Gierl (2007a) is task performance. This model is the most complex in terms of illustrating the cognitive processes that underlie students’ knowledge and skills within a domain. These models include specific details about the thinking processes that students use to solve test items. The authors found that few tests exist that use task performance models as a basis for development. However, they discovered that the development processes undertaken to create cognitively diagnostic assessments (Nichols, 1994) must use the cognitive model
of task performance to be classified as such. Examples of approaches to assessment design that include the cognitive model of task performance as a consideration include ECD, Embretson’s (1994, 1998, 2002, 2005) Cognitive Design System and other diagnostic classification models (see Rupp, Templin, & Henson, 2010). All three processes require identification of the specific knowledge and skills necessary to respond to a test item. Hypotheses about cognitive processes help test developers to better structure their inferential arguments about students’ strengths and weaknesses.

Given that the cognitive model of task performance represented the top of the hierarchy, Leighton and Gierl (2007a) strongly supported its use in test development whenever possible. However, they acknowledge that student thinking processes must be comprehensively detailed, limiting the content focus and extending the time necessary for test development and administration. Still, tests developed from this model are more defensible than any other, assuming the model is empirically tested with a specific student population and alternative task performance models are explored for goodness-of-fit before selecting the model that best predicts student performance. If these criteria are met, “claims about examinees’ thinking processes are unlikely to be attributable to other models” (p. 11). Leighton (2004) advocates for the cognitive model as the basis of test development by stating that these models are validated with the actual knowledge and skills students use when responding to test items. She writes, “this model is the type that researchers develop to confirm empirically that students are employing the expected knowledge and skills on the items developed” (p. 8). Lower-level cognitive models do not attend to actual student processes, thus opening to criticism any inferences made from the results.
Relationships between Types of Cognitive Models. Many domain mastery models are dependent on models of test specifications that preceded them and cannot be considered independent models. For example, New York State developed test specifications through expert consensus, as detailed in the methods section below. For each form of the examination, test items measuring different performance indicators were selected using these specifications as guidance (NYSED, 2009b). The resulting document, with content information as well as skill requirements, represented a cognitive model of domain mastery for the June 2009 form of the Regents Examination in Integrated Algebra. Yet, because it was generated from the lower-level cognitive model of test specifications, the domain mastery model cannot be considered an independent model (i.e., it is dependent on the test specification model). However, Leighton (2004) does allow that “[t]he model of test specifications does not need to be identical to the model of domain mastery because the latter model is general [(i.e., tied to knowledge and skills rather than interpretations)], whereas the model of test specifications is specific to a particular assessment” (p. 7).

This same dependent relationship extends to the most complex model, that of task performance, if either lower-level model is used for generation purposes. However, methods for generating cognitive models of task performance are more often based on theoretical content processing models (Embretson, 1998; Embretson & Gorin, 2001; Svetina et al., 2011) or empirical findings of student protocol analysis (Bonner, 2005; Bonner & D’Agostino, 2012; Briggs & Alonzo, 2009; Campbell, 1999; Gierl, Wang, & Zhou, 2008; Leighton, Cui, & Cor, 2009; Paulsen & Levine, 1999; Roberts, Alves,
Gotzmann, & Gierl, 2009). Leighton, Cui, and Cor (2009) referred to these two different approaches as top-down and bottom-up, respectively.

Because only one model type typically underlies any single test (Leighton, 2004; NRC, 2001), dependent relationships between multi-level models have not received much attention in the literature. Any refinements to a model that lead to a nested structure would signal dependence. However, no studies that specifically examine multiple levels of cognitive models for the same test are found in the literature. One reason for this might be that these models purport to represent the same construct, so repetition is unnecessary. However, given the usefulness of each model type for understanding students’ test performance and making inferences from their scores, the study of multiple cognitive models for one test across all three levels of model complexity has the potential to yield information useful to test validation, test utility, and diagnostic capabilities. Through this study, I aimed to contribute to that effort by comparing an expert-generated domain mastery model, created by enhancing the NYS test specifications, to a student-based cognitive model of task performance for the same assessment.

**Cognitive Models as Validity Evidence**

There is widespread recognition of the need for cognitive models in large-scale assessment (Huff & Goodman, 2007) and many authors have indicated the role that these models could play in validation efforts (e.g., Baxter & Glaser, 1998; Borsboom & Mellenbergh, 2007; Leighton & Gierl, 2011; NRC, 2001; Yang & Embretson, 2007). Leighton and Gierl (2011) have argued intensely for the inclusion of cognitive models in validity arguments for large-scale tests assessments and a few examples are emerging (see Partnership for Assessment of College and Careers (PARCC), 2013; Hendrickson,
Huff, & Luecht, 2010; Kaliski, France, Huff, & Thurber, 2011; O’Callaghan, Morley, & Schwartz, 2004; VanderVeen, 2004). Rodriguez and Haladyna (2013) outlined a course of research to investigate the cognitive behaviors behind selected-response items with a specific focus on gathering validity evidence. Ferrara and DeMauro (2006) reported that validation of large-scale assessments focus on the alignment of content to chosen standards, often based on expert judgments. They went on to state that their review of validity reports lacked evidence to “suggest that test designs are intended to support inferences about student achievement in relation to models of cognition and learning” (p. 613).

Much of the insight into student cognition elicited by test items comes in the form of educated judgments of content experts (Leighton & Gierl, 2007a) and these judgments are rarely, if ever, empirically tested. These findings are not in line with the charge set forth by the NRC (2001) regarding cognitive model requirements. Leighton and Gierl (2011) suggested that this is due to the difficulty of identifying or creating cognitive models for large-scale assessments and issues with the adaptability of such models to a validity argument. Whatever the case, differences between expected and actual student thought processes elicited by a test may raise questions regarding the inferences made from that test (Leighton & Gierl, 2007a).

The difficulty in including models of cognition in validity arguments was met with the conceptualization of the validation of inferences from test scores in terms of students’ possession of attributes. Gierl (2007) operationally defined the term attribute as “[referring] to any procedures, skills, or processes that an examinee must possess to solve a test item” (p. 327). In his seminal chapter on validity, Messick (1989) offered three
different views of content-related evidence for validation. The third of these views cited Lennon (1956) and proposed an examination into the processes used by students to select responses followed by analysis regarding how these processes compare to those typically employed within the content domain. Borsboom and Mellenbergh (2007) used this concept as a foundation to define validity in terms of attributes: “a test is valid for measuring a theoretical attribute if and only if variation in the attribute causes variation in the measurement outcome through the response processes that the test elicits” (p. 93).

The authors elaborated on the assumptions underlying this definition, including the heterogeneity of response patterns for the tested population but the homogeneity of response patterns within each developmental stage. They concluded that the attribute in question moderates the elicited student response processes.

Borsboom and Mellenbergh’s (2007) conception of validation took into account the unique attributes of assessments based on cognitive models. Specifically, the importance of the processes elicited by test items and their relationship with possible interpretations of test results now played a central role in establishing evidence for assessment validity. This evidence was line with the Standards for Educational and Psychological Testing guideline for test use or score interpretations that are dependent upon hypotheses about cognitive processes of examinees (Standard 1.8; AERA, APA, & NCME, 1999, p. 19). This standard required that empirical evidence supporting any such hypotheses be included in the rationale for test selection or interpretation. Additionally, the Standards called for similar evidence to be provided if evidence concerning the cognitive processes of examinees was included in validation efforts.
Borsboom and Mellenbergh (2007) shied away from the widely accepted unified theory of validity (Messick, 1989) but Yang and Embretson (2007) embraced the heading of construct validity and the six aspects under it: content, substantive, structural, generalizability, external, and consequential. The authors specifically focused on the substantive aspect, claiming that it was within this realm that interpretations of examinee performance are made. Evidence to support these interpretations comes in the form of cognitive processes students use to solve test items. This approach reflects the argument-based approach to validation suggested by Cronbach (1982) and House (1980) and more recently extended by Kane (2006, 2013).

Kane (2013) calls for an interpretation/use argument (IUA) and a validity argument that are intertwined rather than sequential. The evaluation of the claims made in the IUA meets the requirement that any claims about the cognitive processes of examinees that affect test interpretation must be supported by empirical evidence (Standard 1.8; AERA, APA, & NCME, 1999, p. 19). To simply assume that experts can accurately predict student processes and, thus, fail to evaluate this claim upon which test interpretation may hinge is referred to as the “begging-the-question fallacy” (Kane, 2013; Walton, 1989). Each claim in the IUA must be supported with evidence in the validity argument.

Kane (2013) outlined the steps of argument-based validation as the process where developers “state the claims that are being made in the proposed interpretation or use, and second, evaluate these claims” (p. 9). In the current study, the interpretations to be made from a test are directly related to the expert-expected processes. However, these interpretations must be supported by evidence that requires examination of the actual
student processes. Yang and Embretson (2007) referred to this evidence as the “construct representation” (p. 122). As cognitive models refer to theoretical constructs (e.g., algebraic problem-solving), an observable indicator is chosen and operationally defined to support specific claims about the construct. Evidence for the appropriateness of indicator selection and definition must be included in the validity argument, along with evidence to support claims about the construct itself (Kane, 2013). For the current study, the indicator was the cognitive processes that content experts expect students to use when solving problems. In the IUA, these processes are claimed to represent the construct measured by the Integrated Algebra examination. In this study, I gathered evidence to support or dispute the indicator (see Research Question 4) as well as the claim about the accuracy of expert-expected processes (i.e., performance indicators; see Research Question 3). Shavelson (2013) referred to this evidence as “cognitive validity” (p. 81), which was the term suggested by Baxter and Glaser (1998) a decade and half earlier.

Incorporating principles for cognitive psychology into educational measurement continues to show promise for establishing construct validity and helping to improve the quality of ability and achievement tests (Embretson & Gorin, 2001; Snow & Lohman, 1989). Baxter and Glaser (1998) demonstrated the use of student cognitive processes in test validation for science assessments. These authors cited Messick’s (1994) condition that task complexity should match the assessed construct and be in line with the level of expertise of the students for whom the test is intended. Baxter and Glaser used think-aloud protocols to elicit the cognitive processes employed by students while solving science items. They found that “the correspondence between the intentions of developers and the nature of elicited performance showed wide variation” (p. 41). It is these
mismatches of cognitive processes that led Leighton and Gierl (2007) to warn about threats to the validity of inferences made from an assessment. Cui and Leighton (2009) suggested the use of cognitive models in test design and validation to strengthen the validity argument with confirmation of students’ cognitive processes underlying test performance. Alves (2012) noted that this approach allows stakeholders to “connect test performance to richer test score interpretations” (p. 20).

**Developing Cognitive Models**

The need to generate cognitive models, whether from theory or empirical investigations, stems from the simple fact that higher-level (i.e., task performance) cognitive models for assessment development mostly do not exist (Leighton & Gierl, 2011). In the past several years, researchers have turned their attention to developing and validating models of cognition that can be used for assessment design purposes (see Leighton & Gierl, 2011 for an extended discussion) but many learning scientists continue to focus on the cognitive processes behind student learning, reasoning, and other thinking-based practices, feeling that this focus is more fundamental than the processes elicited during assessment (e.g., Kuhn, 2001; Lehrer & Schauble, 2006).

To date, the majority of large-scale assessments continue to use Bloom’s taxonomy (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956) to define knowledge and skills (Leighton & Gierl, 2011). The problem with this reliance on cognitive taxonomies such as Bloom’s lies in the lack of credible validity evidence (Schmeiser & Welch, 2006). Leighton and Gierl (2011) insisted that Bloom’s Taxonomy may be cognitive in content, but as it lacks empirical evidence to support the inclusion and placement of
skills, it cannot be considered a cognitive model of learning. Additionally, the continuum of complexity within the Taxonomy has not been verified.

Gorin (2006) suggested that theory is the logical starting point for developing a cognitive model. She pointed out that “theories of cognition, learning, expertise, training, and assessment in various domains can provide rich sources of information for model development” (p. 22). Leighton, Cui, and Cor (2009) agreed with the utility of this approach and added that asking content experts to “review learning outcomes and curriculum objectives, anticipate the relevant knowledge and skills with which to describe a construct, and in their design of test items” (p. 6). This approach, termed expert analysis by the authors, begins with an a priori theory of what knowledge and skills underlie test performance and uses the experience and abilities of experts as evidence of model credibility. They point out that this type of model generation is more common in the field of human information processing than in the realm of educational assessment and testing.

One type of analysis that can be based on expert or student input that has received a great deal of attention in the field of education is the topic of learning progressions. In the short time since the topic first appeared in the literature, learning progressions have taken on a variety of definitional characteristics. Heritage (2013) notes that there are commonalities that unite most of the ways the term has been used. She determined that: “(1) progressions lay out in successive steps, increasingly more sophisticated understandings of core concepts and principles in a domain and (2) progressions describe typical development over an extended period of time” (p. 188). The mapping of student knowledge, skills, and abilities in an ascending manner that demonstrates how students
move from beginner to expert status within a defined domain certainly represents a type of cognitive model. The hierarchical structure is meant to “depict how people likely acquire domain knowledge, while the dependencies, or cognitive connections, among the attributes are clearly indicated” (Broaddus, 2012, p. 7).

Learning progressions are closely related to the attribute hierarchies defined by Leighton, Gierl, and Hunka (2004). Research on learning progressions has shown the usefulness for this type of cognitive model in diagnostic assessment (Briggs, Alonzo, Schwab, & Wilson, 2006; Steedle & Shavelson, 2009). Such models are also accessible to classroom teachers who could employ learning progressions to prepare for large-scale assessments, to design formative assessment for monitoring students’ academic growth, and for information regarding the logical sequence of instruction for a topic (Corcoran, Mosher, & Rogat, 2009; Songer, Kelcey, & Gotwals, 2009).

Many times, theories that might be used to describe student cognition underlying a test are difficult if not impossible to locate. In these cases, developers must look to other avenues for information on student thought processes. Collection of students’ verbalized thoughts while solving test items is useful in gaining an understanding of the skills within the content domain. Indeed, this approach has been used in developing learning progressions. Leighton et al. (2009) identified this process as a bottom-up approach to model generation that requires induction and exploration. Analysis of this think-aloud data has been used to generate cognitive models of task performance in multiple studies (Gierl et al., 2008; Leighton et al., 2009; Roberts et al., 2009). Patterns identified in student responses are organized schematically and become the model, illustrating the cognitive processes of students responding to test items. Although a
growing body of research exists on learning progressions, at the time of this study there was no progression for integrated algebra that could inform the generation of a task performance model.

Think-aloud protocols have been shown to be effective in eliciting the specific knowledge and skills students use when solving problems (i.e., their task performance) (Leighton, 2004; see Ericsson & Simon, 1993; Johnson-Laird, 1983; Newell & Simon, 1972; Siegler, 1994, 1996; Siegler & Crowley, 1991). These reports are important to the development of cognitive models of task performance because they illustrate the actual processes used by students to approach test items (Gallagher, 1992; Gallagher, Levin, & Cahalan, 2002; Gierl, 1997; Hamilton, Nussbaum, & Snow, 1997; Katz, Bennett, & Berger, 2000). Think-aloud protocols can be conducted concurrently or retrospectively to the specific task that participants are undertaking. Concurrent protocols ask participants to verbalize their thought processes while engaging in the task while retrospective protocols require them to recall the steps that they took after completing the task. In their seminal work Ericsson and Simon (1993) review and refute several challenges that arise with the use of think-aloud protocols as a means of data collection for building cognitive models. These issues include is the need for cognitive psychology experts to be involved because educational measurement specialists are not trained in applying cognitive methods, the time and resources necessary for collecting quality cognitive data, and, perhaps most often cited, the questionable trustworthiness of think-aloud protocols. While Ericsson and Simon agree that collaboration with cognitive psychologists is beneficial and that collecting cognitive data is a laborious task, they conclude that student
think-aloud protocols represent a primary source when studying problem solving processes.

The issue of trustworthiness was raised by Nisbett and Wilson (1977), who reported that study participants verbalize what they believe they are doing rather than an accurate portrayal of the cognitive processes they actually employ. Leighton (2004) referred to these reports as “naïve theories” (p. 10). Ericsson and Simon (1980) responded to this issue with the suggestion that think-aloud protocols may be considered a “valuable and thoroughly reliable source of information about cognitive processes” under certain, carefully planned circumstances (p. 247). Such circumstances include thoughtful elicitation of cognitive processes as well as conscientious interpretation of data that takes the context and circumstances surrounding collection into account.

Ericsson and Simon (1980) reviewed additional studies over the next decade and renewed their call for think-aloud protocols as useful data sources for viewing the way in which students approach tasks. Specifically, they recommended concurrent reports, as students were less likely to report their thinking inaccurately when they were asked to verbalize their thoughts as they were solving problems. Fox, Ericsson, and Best (2011) investigated this viewpoint empirically and concluded that think-aloud procedures were non-reactive (i.e., did not change the thought processes being described).

Another issue for consideration when generating a cognitive model of task performance is the point at which it is developed. For cognitively diagnostic assessments, cognitive models of task performance are identified or created a priori and the developers use the model to form the assessment. Items are written carefully so that student responses, whether correct or incorrect, provide diagnostic information in the form of a
demonstration of student skills and understanding (Leighton & Gierl, 2007b; Nichols, 1994). Gorin (2007) suggested that properly written diagnostic items provide information on “why students responded as they did” (p. 174). As noted previously, many large-scale achievement tests are developed from less cognitively complex cognitive models such as models of test specifications or domain mastery. The feedback available from student responses to these items is notably lessened when tasks are not specified (Leighton & Gierl, 2007a). However, it is possible to generate a model of task performance for an existing assessment that was developed under a less complex model.

**Retrofitting Cognitive Models.** In the context of cognitive modeling for educational assessment, retrofitting is the process of generating a cognitive model post hoc from an existing test (Gierl, Alves, & Majeau, 2010). The current study will use retrofitting procedures for the cognitive model of task performance, since the test already exists. Gierl, Wang, and Zhou (2008) retrofitted a cognitive model to mathematics items from the 2005 SAT. The model was developed by experts, validated by having students verbalize their thought processes while solving the items, and compared to a sample of 5,000 student responses. In this case, the model was deemed well-fitting with a Hierarchy Consistency Index (HCI; Cui, Leighton, Gierl, & Hunka, 2006) of 0.80. The HCI is a person-fit statistic that measures the degree of consistency between examinee responses and the specified attribute model. Thus, it appears that in this instance retrofitting was an acceptable method for increasing the granularity of the diagnostic inferences that can be made from an assessment.

Gierl and Cui (2008) warned that applications of retrofitting may be unsuccessful because of the alignment necessary between cognitive models and assessment
components. This lack of alignment diminishes stakeholders’ abilities to make inferences from student responses and scores. Principled test design (e.g., ECD), including a priori model identification, is the preferred method of development when introducing a new assessment (if resource limitations do not prohibit this approach) because alignment between the underlying cognitive model and the components of the assessment instrument can be more purposefully aligned (Gierl et al., 2010). Gierl, Alves, Roberts, and Gotzmann (2009) attempted to retrofit a cognitive model of task performance developed by experts to a group of mathematics items selected from the 2005 and 2006 Preliminary SAT/National Merit Scholarship Qualifying Test (PSAT/NMSQ). They found alignment between their model and the test items of 55% for 2005 and 79% for 2006. Using actual student processes in an attempt to validate their model, they concluded that the model was a poor fit and that diagnostic inferences would be limited and not credible.

The starkly different outcome when models of task performance were created by content experts as opposed to students further underscores the importance of using actual student response processes when developing cognitive models of task performance, especially when a retrofitting approach is selected.

**Cognitive Models in Mathematics**

Given the need for investigation into more complex mathematics and the emphasis by stakeholders such as the National Council of Teachers of Mathematics (NCTM; 2000), the National Mathematics Advisory Panel (NMAP; 2008), and the Common Core State Standards Initiative (CCSSI; 2010), algebra is commonly the focus of studies of cognition and the development of cognitive models. NMAP called for
research into “mechanisms of learning” [mathematics] as well as “features that improve the assessment of mathematical knowledge” (p. xxvi). The organization highlighted algebra skills as crucial to future success in mathematics. Leighton and Gierl (2011) read this call as the organization’s desire for research on empirically-based cognitive models in mathematics. The CCSSI (2010) reports that assessments should reflect not only the content but also the practices of mathematics to prepare students for college and careers.

Mathematics has been at the forefront of the case for using cognitive models in assessment development (e.g., Embretson, 1995; Falmagne & Doignon, 1988; Gierl et al., 2008; Koppen & Doignon, 1990; Leighton et al., 2009; Roberts et al., 2009; Tatsuoka, 1985, 1995; Tatsuoka, Corter, & Tatsuoka, 2004). Likely this is due to the wealth of information available regarding student mathematical reasoning and cognition. Leighton and Gierl (2011) and Alves (2012) each collected a variety of empirical studies found in the learning sciences that guide cognitive model development in mathematics. These include information processing models (e.g., Anderson, 1990; Anderson, Bothell, Byrne, Douglass, Lebière, & Qin, 2004), framework models that identify cognitive processes necessary for mathematical learning (e.g., Kilpatrick, Swafford, & Findell, 2001; Mayer, 2003; 2008), and those types of models originally described by Leighton (2004; e.g., Gierl, Leighton, Wang, Zhou, Gokiert, & Tan, 2009; Leighton, Cui, & Cor, 2009; Tatsuoka, Corter, & Tatsuoka, 2004). However the authors did note that many of these studies are narrowly focused on singular groups and mathematical processes. Additionally, many include only younger children and, thus, early stages of mathematical cognition. This severely diminishes the utility of these investigations to inform assessment. Leighton and Gierl (2011) insisted that “[w]hat test specialists need are
research findings from later stages of mathematical cognition – that is, information about the accumulation of basic processes and how they work together as they materialize into reasoning skills for solving math problems typically encountered by secondary level students” (p. 159). A more detailed examination of the focus of each cognitive study of mathematics as well as other studies that were not directly related to the current investigation can be found in Alves (2012).

**Cognitive Models of Task Performance in Mathematics.** Two distinct approaches to modeling student mathematical cognition are applicable to the current investigation: 1) generating student-based models to be used as models of task performance and 2) comparing expert-generated and student-generated cognitive models. Gierl, Leighton et al. (2009) and Gierl, Wang et al. (2007), when viewed together as a course of research, model student cognitive processes on SAT mathematics items to create a task performance model for interpretation purposes. Leighton et al. (2009) compare an expert-based cognitive model of SAT mathematics with a student-based model and examines the appropriateness of each model for predicting examinee performance on the operational test. These studies provided a basis for the methodological approach of my investigation. Specifically, the work of Leighton and colleagues informed this study by comparing student- and expert-based task performance cognitive models. This study compared an expert-based domain-mastery cognitive model with a student-based cognitive model of task performance.

Gierl, Leighton et al. (2009) engaged 21 high school students in think-aloud protocols to examine their mathematical cognition when asked to solve 21 items taken from the SAT. The researchers first selected items to create a subtest of SAT Algebra I
and II items to be administered to participants. Think-aloud protocols were used to collect data on student cognitive processes as they solved each item. Once data were collected, the researchers created flowcharts to map students’ cognitive procedures as coded from interview audiotapes. The purposes of this step were to graphically present the overall structure of the components and to highlight, compare, and contrast individual differences in elementary steps and solution strategies for each student.

Once student processes were mapped using flowcharts, content experts organized the elicited student processes into hierarchies for each item. That is, individual cognitive processes necessary to solve the item were combined to create a structure with the most complex process or processes at the top and the least complex at the bottom. Each structure represented a cognitive model of task performance for the selected test item and each process was labeled as an attribute.

Once the task performance models were complete, the researchers employed the Attribute Hierarchy Method (AHM; Leighton, Gierl, & Hunka, 2004) for psychometric analysis on a sample of student data. This method classifies students’ responses to test items into “structured attribute patterns associated with different components from a cognitive model of task performance” (Gierl et al., 2009, p. 1). This approach allows for diagnostic inferences to be made from student responses.

Gierl, Leighton et al. (2009) concluded that the models created from the student think-aloud protocols were sufficiently representative of the cognitive processes undertaken to solve the test items. Additionally, they were able to validate the task performance models created by content experts (see Gierl, Wang et al., 2007) by comparing the expert models to student models. Although the expert model fit was good,
the researchers reported that one structural modification to the task performance models was necessary when students’ cognitive processes were examined via think-aloud protocols.

Leighton et al. (2009) also created a cognitive model of task performance for SAT algebra items. In actuality, the researchers created two models: one using the same methods for development with student-reported cognitive processes and one based on expert predictions of students’ knowledge and skills needed to solve the SAT items. The student model was created using think-aloud protocols using the process described by Gierl, Leighton et al. (2009). However, in this case, the researchers sought to develop a cognitive model for algebra as a domain represented by the items. Specifically, the goal was to develop a single model that outlined student cognitive processes for SAT algebra rather than individual models for every item. The expert-based model was created by a expert in the content who reviewed the items selected for the study and identified the cognitive processes that she believed students of any achievement level would need to employ to correctly solve each item. The expert was only able to identify models that applied to all students for eight out of the twenty-one selected items. The researchers reported that the items for which single models could not be identified “reflected a greater diversity of knowledge and skills that required unique models to be specified” (p. 12). Typically, this indicated that students of different achievement levels might approach the item with different processes.

Once the content expert created the task performance model of algebra performance, the student model and expert model were each compared to operational student response data. The expert model was a good fit for the 5,000 randomly sampled
students (mean HCI = 0.71). The student model proved to be a poor fit for this group (mean HCI = 0.48). However, when a subgroup of 100 moderately-high-achieving students was selected, both the student and expert models provided a reasonably good fit (student mean HCI = 0.93, expert mean HCI = 0.74) demonstrating that both models predicted moderate to high-achieving student performance in an operational test setting with some level of accuracy. The authors concluded that these results “suggested that both expert analysis and verbal analysis were useful for generating cognitive models to account for examinee data,” yet the predictive ability of the student model is notably superior for moderate to high-achieving students (Leighton et al., 2009, p. 246). The authors did not compare the models qualitatively or seek to identify any differences between the cognitive models and the construct measured by the test. Their analysis was limited to regression-based predictability measures.

The resultant improved model fit for moderately-high-achieving students questioned the assumption that one model is sufficient for all students. Underlying this assumption is the concept that the performance of high- and low-achieving students is “distinguished simply by the mere presence or absence of attributes within a single model” (Leighton et al., 2009, p. 248). The researchers’ results suggested that this may be linked to differences in problem solving ability attributed to cognitive development (see Chi, Glaser, & Farr, 1988; Mislevy, 2006; Sternberg, 1999).

The two studies described above represent the current research directions of cognitive models of task performance in mathematics. Although both studies generated cognitive models for existing tests (i.e., retrofitting), neither investigation examined specific student processes in any depth. Leighton et al. (2009) generated both expert and
student models, but only used regression-based analyses to explore the prediction abilities of each model source. As both were deemed appropriate for certain student populations, the differences between the models were not highlighted. My study compared each type of model (i.e., expert- and student-generated) to see if differences between the processes expected by experts and the actual processes used to approach test items exist.
Chapter 2: Method

Participants

Ten participants for this study were recruited from secondary schools in the Capital District of New York State. As Integrated Algebra was the content area under investigation, participants had to have completed within the past year, or be nearly finished with, a year-long course in algebra to be eligible to take part in the study. Typically, these students were in eighth, ninth, or tenth grade and approximately 13-16 years old. The sample included six females and four males. Although it remains controversial and difficult to draw any firm conclusions, there exists a body of evidence suggesting that males and females might approach problem solving in math class differently (e.g., DeMars, 1998; Gallagher, 1992; Gallagher, & De Lisi, 1994; Gallagher, De Lisi, Holst, McGillicuddy-De Lisi, Morely, & Cahalan, 2000; Gallagher, Levin, & Cahalan, 2002; Garner & Englehard, 1999; Gierl, Bisanz, Bisanz, & Boughton, 2002). Zhu (2007) reviewed this literature and concluded that “we may assume that males and females have different patterns of mathematical problem solving” (p. 189). These differences were not part of this investigation, but given the research question devoted to validation efforts, it seemed important to note if any such pattern of procedural differences did in fact occur.

There were some necessary qualifications for participation beyond completion of the algebra coursework. Bonner and D’Agostino (2012) encouraged the use of participants that are preparing for, or at least very familiar with, the specific test used to elicit cognitive processes (see instrumentation section below). To best access this population, students who were preparing or just finished with the June 2013 exam were
prioritized. Additionally, students who took the exam in June 2012 were eligible as they would be familiar with the test. Additional evidence that these students should be able to provide details into their cognitive processes relevant to Integrated Algebra comes from the scaffolded nature of New York State Content Standards for Mathematics (NYSED, 2005): The next course in the sequence of mathematics is Geometry, which builds on many of the concepts necessary for success in algebra. The stepwise nature of the coursework suggested that students would retain much of their algebra knowledge and abilities up to the point of data collection for this study.

In addition to course and test requirements, Leighton, Cui, and Cor (2009) recommended several constraints to “narrow the pool of potential participants” (p. 241). It is crucial that participants are at an achievement level high enough to provide a high proportion of correct item responses and detailed verbalization of their cognitive processes. Correctness and ability to verbalize are necessary to understand the cognitive processes used to solve problems on the exam. While incorrect steps do not preclude development of cognitive models, the cognitive model underlying a test should model correct responses (i.e., those that allow students to demonstrate mastery of the construct). Thus, students likely to respond correctly to test items are of greater use for model generation. Additionally, lower-achieving students may take unnecessary or incorrect procedural steps and experience frustration or an inability to explain their thought processes (Kutetskii, 1976; Leighton et al., 2007). Either of these situations would be detrimental to data collection because they would contribute processes irrelevant to solving the test items.
Another concern is that very high-achieving students might automatically solve test items without conscious thought to the cognitive steps taken. This automaticity would provide little or no data useful in creating a cognitive model (Ericsson & Simon, 1993; Leighton, 2004). Given the above constraints, students at moderate to moderately-high achievement levels, as determined by score on the NYS Regents Examination in Integrated Algebra or prior performance in mathematics courses, were selected for participation in this study. This information was collected only anecdotally through the parents and students. For example, students were asked if they liked math and if they scored well on in the class or on the Regents Exam. Students who responded affirmatively were invited to participate.

There is much debate surrounding the appropriate sample size for data collection on problem solving processes via concurrent think-aloud protocols. A review of studies employing similar methodology revealed a range in sample size from five to 49 students. Almond et al. (2009) propose that one key to determining an appropriate sample size lies in labeling the study as exploratory or confirmatory. This study fell in between the two classifications because one cognitive model (i.e., that of domain mastery and expert beliefs about student processes) is under confirmatory review but the nature of the data collection procedure would suggest an exploratory study. To bridge this divide and ensure the sample was appropriate for either classification, the sample for this study included 10 participants. This sample size did not fully meet the more stringent requirement of confirmatory studies, but increased the likelihood of saturation (Almond et al., 2009; Creswell, 2007; Glaser & Strauss, 1967). This sample size was sufficient to provide robust insight into the cognitive processes elicited by the items.
Instrumentation

June 2009 Regents Examination in Integrated Algebra. The test items used in this study came from the New York State Regents High School Examination in Integrated Algebra – June 2009, which is available to the public in electronic format (nysedregents.org). This particular version of the test was chosen as the instrument for this study because it is the most recent test for which full technical reports, including the test development report and the reliability and validity report, were available from the New York State Education Department (NYSED). The Integrated Algebra exam is the first in a sequence of three tests (Integrated Algebra, Geometry, and Algebra 2/Trigonometry) created to reflect the NYS content standards for mathematics (NYSED, 2005). The New York State Education Department “requires students to earn three units of credit in high school mathematics and pass, with a 65 or higher, one Regents Examination in mathematics” (NYSED, 2009b, p. 3). Typically, the Integrated Algebra exam is taken to fulfill the examination requirement as it is the first test in the series of three. The results of the examination are also used for accountability purposes under the No Child Left Behind Act of 2001 (NCLB, 2002; NYSED, 2009b).

According to NYSED (2009b), all Regents Examinations, including Integrated Algebra, are designed to:

- Evaluate higher-order thinking skills and performance abilities, including planning and acquiring resources, designing and problem solving, conducting independent research, and producing real-world products.
- Provide information that helps teachers adapt instruction to students’ strengths and needs and that informs students, parents, educators, and the general public about what students are expected to know and be able to do (p. 9).
This design information is helpful in classifying the purposes of the test and demonstrating that it is appropriate to expect an underlying cognitive model. If the test can successfully measure higher-order thinking skills and performance abilities, it is crucial that these traits are present in the cognitive model and that actual student response processes reflect them.

Students are classified into one of three achievement levels based on their performance on the examination. The achievement levels and corresponding scale scores are: Level I: (0-64), Level II (65-84), and Level III (85-100). A scale score of 65 must be earned to pass the exam, which consists of 39 items: 30 multiple-choice items each worth two credits, three 2-credit constructed-response items, three 3-credit constructed-response items, and three 4-credit constructed-response items for a total of 87 credits. After adding up earned credits for a total raw score, a conversion chart is used to place the raw score on a 0-100 scale for student, teacher, and parent interpretation. Although meant to show if a student has met the passing cut score of 65, the 100-point scale often causes confusion as it is incorrectly interpreted as representing percentage of correct responses or student percentile rank.

The Office of State Assessment at NYSED engages in a multistep process for assessing the reliability and validity of Regents examinations. Evidence for the reliability of the June 2009 Integrated Algebra exam was sought through measures of internal consistency, standard error of measurement, and classification accuracy. All of these analyses, taken together, provide evidence for the reliability of the exam. Cronbach’s (1951) alpha, a measure of internal consistency, was 0.93 for the full test and 0.88 for the multiple-choice section only. This measure reports a high level of internal consistency.
The reliability estimates were calculated for all students and various demographic subgroups. For each of these subgroups, the reliability estimates (i.e., Cronbach’s alpha) ranged from 0.86 to 0.93 which, again, is reasonably high for a test of this type (NYSED, 2005). Finally, analysis of classification (Expected Classification Accuracy; Rudner, 2005) accuracy demonstrated that 89.6% of students were classified into the proper performance level. At the cut score of 65, the false positive rate was 3.9% and the false negative rate was 1.9%.

NYSED currently undertakes a rigorous validation process that compiles evidence before, during, and after development and administration of the test. Content validity evidence is heavily geared toward expert input into all aspects of test design. Experts who are licensed educators write, revise, and select items for inclusion on the operational test under the guidance of a test owner (NYSED, 2009c).

Evidence for construct validity includes item-total correlations (i.e., item discrimination), Rasch fit statistics, correlation among content strands, correlation among item types, principal component analysis and differential item functioning analyses to search for differences in performance between demographic subgroups. Point-biserial (i.e., item-total) correlations ranged from 0.23 to 0.70 for all test items and from 0.23 to 0.58 for multiple-choice items only. These values indicate that the items adequately differentiated between high- and low-achieving students on a common trait. Additional evidence of unidimensionality was seen in the Rasch fit statistics. The mean square fit statistics, in most cases, ranged from 0.7 to 1.3. Correlations among the five content strands (Number Sense and Operations, Algebra, Geometry, Measurement, and Statistics and Probability) were heavily influenced by the number of items within each strand,
which ranged from three to twenty-one. Still, these correlations were between 0.45 and 0.77. Item-type correlations were high as well (0.85 - 0.98). Results of a principal component analysis revealed a dominant factor “accounting for 80% of variance among factors with eigenvalues exceeding 1” (NYSED, 2009c, p. 22). Finally, differential item functioning analyses showed minimal differences across ethnic and gender subgroups. All of this evidence suggests that the June 2009 Regents Examination in Integrated Algebra was measuring a unidimensional construct.

**Items selected for study.** Nine items were selected for inclusion in this study as this number was deemed an approximate ceiling for cognitive interviews by Almond et al. (2009) based on participant fatigue and the likelihood of reaching data saturation. Items were chosen based on content specifications and on parameter estimates obtained from the results of stand-alone field testing procedures. New York State uses an average Rasch item difficulty of 0.083 as a target for all administrations of the Integrated Algebra Regents exam. To be statistically eligible for selection for this study, items needed to fall within approximately one standard deviation of this statistic (i.e., -1 ≤ Θ ≤ 1). Additionally, selected items had acceptably high point-biserial correlations (> 0.30), p-values above 0.40, and showed reasonable variation during distractor analysis (no option < 5%; see Table 1 for selected item statistics). From a content standpoint, the resulting items came from three of the five content strands within Integrated Algebra and measured no duplicate performance indicators. The selected items and their corresponding content strands and performance indicators are listed in Table 2. The full text of these items is presented in Appendix E.
Table 1

*Item Statistics for Selected Items*

<table>
<thead>
<tr>
<th>Item Position</th>
<th>Rasch Item Difficulty Estimate</th>
<th>Weighted Item Mean</th>
<th>Point-Biserial</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>-0.33</td>
<td>0.60</td>
<td>0.47</td>
</tr>
<tr>
<td>5</td>
<td>0.02</td>
<td>0.61</td>
<td>0.43</td>
</tr>
<tr>
<td>9</td>
<td>-0.97</td>
<td>0.70</td>
<td>0.57</td>
</tr>
<tr>
<td>10</td>
<td>-0.37</td>
<td>0.61</td>
<td>0.51</td>
</tr>
<tr>
<td>14</td>
<td>-0.17</td>
<td>0.61</td>
<td>0.39</td>
</tr>
<tr>
<td>15</td>
<td>-0.18</td>
<td>0.57</td>
<td>0.36</td>
</tr>
<tr>
<td>16</td>
<td>0.02</td>
<td>0.57</td>
<td>0.58</td>
</tr>
<tr>
<td>17</td>
<td>-0.07</td>
<td>0.49</td>
<td>0.44</td>
</tr>
<tr>
<td>18</td>
<td>-0.11</td>
<td>0.70</td>
<td>0.46</td>
</tr>
</tbody>
</table>

Table 2

*Content Strands and Performance Indicators for Selected Items*

<table>
<thead>
<tr>
<th>Item Position</th>
<th>Content Strand</th>
<th>Performance Indicator</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Algebra</td>
<td>A.A.27 - Understand and apply the multiplication property of zero to solve quadratic equations with integral coefficients and integral roots</td>
</tr>
<tr>
<td>5</td>
<td>Statistics and Probability</td>
<td>A.S.1 - Categorize data as qualitative or quantitative</td>
</tr>
<tr>
<td>9</td>
<td>Algebra</td>
<td>A.A.45 - Determine the measure of a third side of a right triangle using the Pythagorean theorem, given the lengths of any two sides</td>
</tr>
<tr>
<td>10</td>
<td>Number Sense and Operations</td>
<td>A.N.2 - Simplify radical terms (no variable in the radicand)</td>
</tr>
<tr>
<td>14</td>
<td>Algebra</td>
<td>A.A.21 - Determine whether a given value is a solution to a given linear equation in one variable or linear inequality in one variable</td>
</tr>
<tr>
<td>15</td>
<td>Statistics and Probability</td>
<td>A.S.9 - Analyze and interpret a frequency distribution table or histogram, a cumulative frequency distribution table or histogram, or a box-and-whisker plot</td>
</tr>
<tr>
<td>16</td>
<td>Algebra</td>
<td>A.A.15 - Find values of a variable for which an algebraic fraction is undefined.</td>
</tr>
<tr>
<td>17</td>
<td>Algebra</td>
<td>A.A.6 - Analyze and solve verbal problems whose solution requires solving a linear equation in one variable or linear inequality in one variable.</td>
</tr>
<tr>
<td>18</td>
<td>Algebra</td>
<td>A.A.41 - Determine the vertex and axis of symmetry of a parabola, given its equation (See A.G.10)</td>
</tr>
</tbody>
</table>
Integrated Algebra Test Specifications. As reported in Chapter 2, a test specifications model represents the lowest level of cognitive model (Leighton & Gierl, 2007) and provides the basis of any higher-level cognitive models that are generated. The NYS Education Department uses a table of specifications in the design of all Regents examinations. The technical report for the June 2009 Regents examination in Integrated Algebra describes the process undertaken to develop these specifications:

A meeting was held in November 2006 with sixty-three professional New York State educators to determine the test specifications for the Regents Examination in Integrated Algebra. The purpose of these specifications is to document the necessary requirements for item types and the emphasis per content strand. The method used for determining the test specifications was to divide the educators into two groups that made independent recommendations for the test specifications and then came together to agree on a final recommendation that was sent to the New York State Education Department (NYSED). The NYSED considered the recommendation, along with other factors, and provided a final decision on the Integrated Algebra Test Specifications (NYSED, 2009b, p. 11).

The resulting specifications (see Table 3) represent a general guide for the experts to follow when writing and selecting items for the operational exam. However, a much more detailed document, specific to the particular test form, is created following form construction. This document outlines the item position, item type, maximum credits, content strand, and performance indicator for each item on the specific operational test and is closer in attributes to Leighton and Gierl’s (2007) concept of a domain-mastery cognitive model (DMCM). The DMCM for the June 2009 Regents Examination in Integrated Algebra is presented as Table 4. This is the expert-based cognitive model of domain mastery that was compared to the task performance cognitive models (TPCMs) that were developed as a part of this investigation.
Table 3

*Regents Examination in Integrated Algebra Test Specifications (NYSED, 2009b, p. 12)*

<table>
<thead>
<tr>
<th>Content Strand</th>
<th>Percentage of Total Credits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number Sense and Operations</td>
<td>6 – 10%</td>
</tr>
<tr>
<td>Algebra</td>
<td>50 – 55%</td>
</tr>
<tr>
<td>Geometry</td>
<td>14 – 19%</td>
</tr>
<tr>
<td>Measurement</td>
<td>3 – 8%</td>
</tr>
<tr>
<td>Statistics and Probability</td>
<td>14 – 19%</td>
</tr>
</tbody>
</table>

Table 4

*June 2009 Regents Examination in Integrated Algebra Cognitive Model of Domain*

*Mastery (NYSED, 2009b, 2009d)*

<table>
<thead>
<tr>
<th>Item Position</th>
<th>Item Type</th>
<th>Maximum Credits</th>
<th>Content Strand</th>
<th>Performance Indicator</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Measurement</td>
<td>A.M.1</td>
</tr>
<tr>
<td>2</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Algebra</td>
<td>A.A.27</td>
</tr>
<tr>
<td>3</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Algebra</td>
<td>A.A.12</td>
</tr>
<tr>
<td>4</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Algebra</td>
<td>A.A.1</td>
</tr>
<tr>
<td>5</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Statistics and Probability</td>
<td>A.S.1</td>
</tr>
<tr>
<td>6</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Algebra</td>
<td>A.A.4</td>
</tr>
<tr>
<td>7</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Algebra</td>
<td>A.A.25</td>
</tr>
<tr>
<td>8</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Statistics and Probability</td>
<td>A.S.21</td>
</tr>
<tr>
<td>9</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Algebra</td>
<td>A.A.45</td>
</tr>
<tr>
<td>10</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Number Sense and Operations</td>
<td>A.N.2</td>
</tr>
<tr>
<td>11</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Measurement</td>
<td>A.M.2</td>
</tr>
<tr>
<td>12</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Algebra</td>
<td>A.A.7</td>
</tr>
<tr>
<td>13</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Algebra</td>
<td>A.A.23</td>
</tr>
<tr>
<td>14</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Algebra</td>
<td>A.A.21</td>
</tr>
<tr>
<td>15</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Statistics and Probability</td>
<td>A.S.9</td>
</tr>
<tr>
<td>16</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Algebra</td>
<td>A.A.15</td>
</tr>
<tr>
<td>17</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Algebra</td>
<td>A.A.6</td>
</tr>
<tr>
<td>18</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Algebra</td>
<td>A.A.41</td>
</tr>
<tr>
<td>19</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Geometry</td>
<td>A.G.3</td>
</tr>
<tr>
<td>20</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Geometry</td>
<td>A.G.6</td>
</tr>
<tr>
<td>21</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Algebra</td>
<td>A.A.16</td>
</tr>
<tr>
<td>22</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Algebra</td>
<td>A.A.34</td>
</tr>
<tr>
<td>23</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Algebra</td>
<td>A.A.13</td>
</tr>
<tr>
<td>24</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Geometry</td>
<td>A.G.8</td>
</tr>
</tbody>
</table>
Data Collection

Following the think-aloud protocols, the set of nine Integrated Algebra items were administered to individual students while I observed and took notes. After briefing each subject on think-aloud protocols, I asked participants to verbalize their thinking while they attempted to solve the items. This process was videotaped for later transcription to be supplemented with my notes. All verbalized cognitive processes from students were recorded and participants were encouraged to share any thoughts that they had while solving the items. The emphasis was placed on describing cognitive processes, not on other test-taking variables such as speededness.

Informed parental consent was necessary for participants to be eligible to take part in the study. Once parental consent was obtained, students were presented with the purpose of the study and methods used and were given the opportunity to ask questions or raise concerns. After this, student assent was sought and the student’s comfort with
and understanding of the process and all aspects of the study was confirmed. At this point, I read the standardized instructions in Appendix A to students regarding the procedures for verbalizing their thought processes.

Before beginning the Integrated Algebra item set, I used a warm-up strategy suggested by Almond et al. (2009) that included modeling and a practice item. First, I solved an item and verbalized my thoughts while reading the item, engaging in problem solving, and selecting a response. Next, the student was given an item that was not part of the study set and was informed that this item was a chance to practice thinking-aloud similarly to the model they just witnessed. I encouraged students who were not verbalizing all of their thoughts to think-aloud more often. Additionally, students were able to ask questions or receive reassurance about the appropriateness of their problem solving and verbalization skills (Bonner, 2005). Following the warm-up strategy, I asked the student if he or she was ready to move onto the study item set and from that point on, I no longer involved myself with the think-aloud process except to remind students to verbalize their thoughts.

After students worked through all nine items, they were asked to reflect on how they solved each item (i.e., a retrospective report). This offered the opportunity to clarify any points from the students’ concurrent verbalizations and allowed for the recognition of metacognitive processes that were not revealed during the original think-aloud (Leighton, 2005). Once the retrospective interview concluded, any feedback about the study and data collection process was solicited. After this segment, the interview was terminated.
Data Analysis

Construction of cognitive models of task performance. Student think-aloud protocols were transcribed from the videotaped interviews. The resulting codes are at a fine-grain level of detail because of the need to compare student processes with the performance indicators developed by content experts (Katz, Bennett, & Berger, 2000). Specifically, the coding for fine-grain information requires students’ individual cognitive processes to be coded for each item. Although this level of detail of coding may have been more detailed than necessary for comparison purposes, Bonner (2005) reported that a fine-grain analysis “maximize[s] the amount of information obtained from the protocols about the presence or absence of various kinds of thinking” (p. 56). For validation efforts, it was important to record what cognitive processes may be lacking and if these absences affected students’ ability to respond to test items.

After the think-aloud protocols, I watched the taped interviews and viewed the transcriptions. During coding, I referred to a strategy established to guide the coding process (see Appendix G). This strategy dictated that the coding of cognitive processes began after the student read each item and the processes were coded at the level of individual cognitive processes (i.e., every time the student begins a new mathematical operation). For example, a student solving $2x + 1 = 7$ for $x$ might have used the following steps:

1. Subtract 1 from both sides of the equation
2. Divide both sides of the equation by 2
3. Select option $x = 3$

Although the overarching work may be seen as finding the solution to an algebraic equation, the cognitive processes within this solution were the desired level of coding. An additional component of the strategy was that any additional processes elicited during the
students’ retrospective reports would be included in the solution path because the processes were used while responding to the item; they were just not verbalized at that time.

The coding policy included rules for recognizing when to record student processes as well as how to handle student scenarios such as guessing, changing strategies, or saying “I don’t know.” The context of the cognitive process was crucial to accurate coding because students may struggle with vocabulary (or the think-aloud protocol process), but still use a certain process. As long as the cognitive process was explicit in students’ words or writing, it was included. The full list of policy guidelines is included in Appendix G.

To calibrate the coding, a second rater and I independently coded eighteen (i.e., two flowcharts for each item) randomly selected items into flowcharts (see Appendix G for the process for rater calibration). The second rater was a doctoral candidate in educational psychology and former secondary science teacher. She was comfortable with this level of mathematics and was familiar with the content. Following our independent coding, the rater and I engaged in a discussion about our process of and reasoning for the coding, compared our coding, and identified discrepancies. There were zero discrepancies in terms of coding cognitive processes. The coding required only the identification of individual cognitive processes and the other rater and I did not disagree on any of the processes as we had discussed the coding strategy, coding policy, and appropriate grain size before beginning the rater calibration. However, we did engage in a discussion about terminology and how to best present student cognitive processes in
flowchart form. This discussion was productive and we were able to agree on the terminology that would be used for the cognitive processes within the flowcharts.

Once this rater calibration was completed, I created flowcharts showing the knowledge and skills used by the students to approach each item and ultimately select a response. These knowledge and skills included mathematical processes (e.g., take the square root of 16), general problem-solving tactics (e.g., review each option), or test-taking strategies that skirted the need for content knowledge (e.g., eliminate options not in table of values). This process resulted in 75 flowcharts (10 students X 9 items – 15 incorrect responses). A sample flowchart for item 2 is shown below in Figure 2 and all of the charts are included in the next chapter. Each time I completed coding approximately ten flowcharts, I rechecked calibration by comparing my flowcharts to those coded with the second rater. This helped to ensure the coding fidelity was preserved.
Next, I identified cognitive processes that appeared across different students’ think-aloud protocols for an item. Once common processes employed by students were identified, I coded the think-aloud reports (i.e., transcriptions) using the newly identified common cognitive process. This allowed me to associate multiple students with a single solution path when applicable (i.e., when students used similar cognitive processes in responding to the item). However, if a solution path was unique to a student, it was also included to ensure that no data were ignored. The result was nine cognitive models of task performance (i.e., one for each Integrated Algebra item) specific to the selected sample of participants. Inferences were then made regarding the cognitive processes necessary to solve the items, the actual processes used by the participants, and the relationships between processes and items. For example, an abundant overlap of skills across different items was visible following the creation of the TPCMs. Data saturation
was determined by noting that at least one duplicate solution path emerged for each item, and that there were marked similarities and duplications in student solution paths for each item.

**Comparison of cognitive models of domain mastery and task performance.**

Following the creation of the TPCMs, I compared the cognitive processes used by students to solve the algebra items shown in the TPCM for each item with the performance indicators in the DMCM. Specifically, I checked for alignment between the expected cognitive processes used to solve the items (i.e., performance indicators) and the actual cognitive processes elicited by the think-aloud protocols and expressed in the task performance cognitive model. This analysis took place at the item level and the results of the analysis reveal whether students actually used the processes expected by the content experts developing the test. Frequency analyses of agreement between the DMCM and TPCMs were conducted and incidents of agreement and disagreement between the models were reported for each item.
Chapter 3: Results

In this chapter, I present the results of the think-aloud protocols conducted with students solving multiple-choice integrated algebra items. The chapter is divided into two sections that address the four research questions driving the study:

1. What are the cognitive processes students use to solve nine selected items on the Integrated Algebra Regents examination?

2. How can students’ cognitive processes be represented in cognitive models of task performance?

3. To what extent do the cognitive processes used by students (i.e., the task performance cognitive models) align with the performance indicators expected by content experts for each item (i.e., the domain mastery cognitive model)?

4. To what degree does the construct validity evidence provided by examination of expert-expected and actual student processes support the claims of the exam?

In the first section, I report on the cognitive processes used by students to solve the nine multiple-choice integrated algebra items and present their organization into item-level cognitive models of task performance. I briefly describe the most common strategies that students selected to solve the problems. Additionally, I identify situations where evidence of test-wise strategies appeared in the student think-aloud protocols. Such strategies included elimination of answer options or exploitation of item-writing limitations. In the second section, I offer the results of the comparison between cognitive processes expected by the content experts involved with test development and actual student processes and discuss the evidence that this comparison provides for construct validation.
Research Questions 1 and 2

These research questions addressed what the cognitive processes employed by students were and how these processes might be represented as cognitive models of task performance. The cognitive processes used by students to solve the multiple-choice integrated algebra problems were elicited through the think-aloud protocols. These processes were organized into flowcharts showing the processes used by students who responded correctly to the item. Gierl et al. (2007) referred to the data represented in the flowcharts as solution paths. The processes of students who responded incorrectly are not presented because, while incorrect processes may provide insight into student misconceptions and other diagnostic information (e.g., Luecht, 2006), expert-expected processes anticipate student success. Thus, there is no basis for comparing incorrect student processes with the cognitive model of domain mastery. As anticipated given the target characteristics of the sample, a high percentage of the students responded correctly to each item. Table 5 displays the number of students who responded correctly to each item. This number ranged from seven to ten out of ten total participants. This high rate of success was helpful in ensuring that data saturation was reached.

Table 6 shows the performance of each student on the items. Student success ranged from five to nine out of nine items, or 55.5-100.0% correct. To participate, students had expressed comfort with mathematics as well as previous success. No student expressed concern that they were unfamiliar with the content. Interview comments typically reported a lack of remembering after students struggled with items. Additionally, approximately one-third of the incorrect responses were the result of conceptual or calculation mistakes. Taken together, Tables 5 and 6 show no patterns in
student performance that suggest any particular item was problematic. Additionally, direct comparisons of the solution paths of male and female participants for each item yielded no pattern of differences in cognitive processes between genders, as Zhu (2007) suggested might occur.

Table 5

*Number of Students Responding Correctly to Selected Items*

<table>
<thead>
<tr>
<th>Item #</th>
<th>Number of Correct Responses (out of 10 total students)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>14</td>
<td>8</td>
</tr>
<tr>
<td>15</td>
<td>7</td>
</tr>
<tr>
<td>16</td>
<td>7</td>
</tr>
<tr>
<td>17</td>
<td>8</td>
</tr>
<tr>
<td>18</td>
<td>9</td>
</tr>
</tbody>
</table>

Table 6

*Student Performance on Selected Items*

<table>
<thead>
<tr>
<th>Student</th>
<th>Items Answered Correctly</th>
<th>Item Answered Incorrect</th>
<th>Percent Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5,9,10,15,17</td>
<td>2,14,16,18</td>
<td>55.6</td>
</tr>
<tr>
<td>2</td>
<td>2,5,9,10,14,15,16,18</td>
<td>17</td>
<td>88.9</td>
</tr>
<tr>
<td>3</td>
<td>2,5,9,10,14,15,16,17,18</td>
<td></td>
<td>100.0</td>
</tr>
<tr>
<td>4</td>
<td>2,5,9,14,18</td>
<td>10,15,16,17</td>
<td>55.6</td>
</tr>
<tr>
<td>5</td>
<td>2,5,9,14,16,17,18</td>
<td>10,15</td>
<td>77.8</td>
</tr>
<tr>
<td>6</td>
<td>2,5,9,10,15,16,17,18</td>
<td>14</td>
<td>88.9</td>
</tr>
<tr>
<td>7</td>
<td>2,5,9,10,14,15,16,17,18</td>
<td></td>
<td>100.0</td>
</tr>
<tr>
<td>8</td>
<td>2,5,9,10,14,15,16,17,18</td>
<td></td>
<td>100.0</td>
</tr>
<tr>
<td>9</td>
<td>5,9,10,14,16,17,18</td>
<td>2,15</td>
<td>77.8</td>
</tr>
<tr>
<td>10</td>
<td>2,5,9,10,14,15,17,18</td>
<td>16</td>
<td>88.9</td>
</tr>
</tbody>
</table>

Once the solution paths for each item were identified for each student who responded correctly, all of the flowcharts for a given item were combined. This resulted
in a single flowchart for each item showing multiple solution paths. These flowcharts represent cognitive models of task performance specific to the sample of students who participated in this study. Figures 3-11 present the processes students used to correctly respond to the multiple-choice items as TPCMs (note that Figure 10 appears on two pages due to space constraints).

Each TPCM begins with students reading the item and concludes with the students selecting an option (i.e., multiple-choice response). Each of these is shown in a start/stop box (i.e. an ellipse labeled “Read Problem Aloud”) to be consistent with standard flow chart symbols. The cognitive processes are represented in the rectangular process boxes. Coding occurred at a fine-grain level to capture each separate cognitive process. All solution paths are connected by flowlines. Recorded below the selected option (i.e., stop box) is the participant number of any students that followed that particular solution path.
Figure 3. Cognitive model of task performance for item 2: *What are the roots of the equation* \(x^2 - 7x + 6?\) (NYSED, 2009a).

Figure 3 presents the TPCM for item 2 on the exam. The numbers below each selected option (i.e., stop boxes) reflect which students chose this particular solution path. For example, students 3, 4, 7, and 8 arrived at option 4 for the item by factoring into binomials, setting each binomial equal to zero, and solving each for \(x\). In this item, students were asked to identify the roots of a quadratic equation. Six out of the eight students who responded correctly to this item chose to factor the equation into binomials. While each of these students adopted a similar approach, three solution paths result due to one student opting to check the answer before selecting an option and one student who graphed the equation and identified the roots visually before factoring. One student substituted the answer options into binomials and applied the distributive property in an
attempt to match the equation given in the item stem. Finally, one student guessed based on the constant term in the given equation.

Figure 4. Cognitive model of task performance for item 5: Which data set describes a situation that could be classified as qualitative? (NYSED, 2009a).

The TPCM for item 5 is shown in Figure 4. Given four options, students needed to identify a dataset containing information classified as qualitative. It would seem that this item would invite a test-wise strategy. A test-wise strategy is an approach that focuses on the aspects of the item itself rather than the content with an eye towards finding clues to the correct response. This item requires students to identify a qualitative data set. Presumably, three of the four options share at least one characteristic that is not incorrect while one option, the correct response, does not share the characteristic. Students could employ an elimination strategy that would not require any knowledge.
about the definition of the term *qualitative*. Instead, students in the study stated applicable definitions to set parameters for reviewing the options before actually beginning their analysis. All students responded correctly to this item, differing only on how they defined the term *qualitative*. Seven students defined it as referring to qualities, while three simply identified qualitative as the alternative to quantitative (i.e., not numeric). Students who defined both terms chose to classify each option before selecting the qualitative dataset. Those who viewed the terms relative to each other identified the qualitative option by looking for the option not involving numeric data.

Figure 5. Cognitive model of task performance for item 9: *What is the value of x, in inches, in the right triangle below?* (NYSED, 2009a).
Figure 5 shows the TPCM for item 9, which required students to solve for the length of the hypotenuse of a right triangle, given the other two side lengths. Again, all students responded correctly to this item. This may be due to the fact that the application of the Pythagorean Theorem is introduced in earlier grades and students were likely familiar with the concept before enrolling in Algebra. Another hypothesis is that the item may only require substituting numbers into a provided formula and solving the equation without specifically accessing a deep understanding of the Pythagorean Theorem. The figure shows two solution paths which differ only in how students dealt with the radical number resulting from the Pythagorean Theorem. Six students attempted to reduce the radical, while four other recognized that their result matched the third option for the item.
The TPCM for item 10 is presented in Figure 6. Three solution paths emerged as students simplified a radical number. For two of these paths, students broke the radical number into factors where one of the factors was a perfect square. Then they were able to remove one of the factors from under the radical sign. The third solution path followed a similar strategy, but did not choose the largest perfect-square factor. This required the student to perform the same steps twice in order to arrive at the answer. Still, the concept elicited was the same as the students who executed the simplification in fewer steps.
Figure 7. Cognitive model of task performance for item 14: Which value of $x$ is in the solution set of $\frac{4}{3}x + 5 < 17$? (NYSED, 2009a).

In Figure 7, the TPCM for item 14 is shown. This model shows one of two instances for the selected items where students adopted, and successfully applied, a test-wise strategy other than guessing. The item required students to identify which of four options was in the solution set of a given inequality. Four students approach this by substituting the possible answers into the given algebraic inequality and evaluating the result. This approach was successful in each case. Students adopting a more straightforward mathematical approach isolated the variable in the inequality before examining the answer options for a number that was in agreement with their resulting inequality. One student substituted the option back into the original inequality as a checking method before finalizing her selection.
Figure 8. Cognitive model of task performance for item 15: *The box-and-whisker plot below represents students’ scores on a recent English test. What is the value of the upper quartile?* (NYSED, 2009a).

The TPCM for item 15 is displayed in Figure 8. Item 15 showed a box-and-whisker plot and asked students to identify the upper quartile. As this was an assessment of declarative knowledge, students did not take many steps in selecting an option. Students who responded incorrectly simply lacked the knowledge necessary to identify the upper quartile. Perhaps because such specific knowledge was tested through this item, this was one of the items on which the participants performed the worst and seven out of ten students answered the item correctly. Of those students who selected the correct option, all but one identified each data point shown on the plot before calculating the scale and finding the value of the upper quartile. The student who did not take this route chose to simply identify only the upper quartile on the plot, calculate the scale, and match his or her identification to the available options.
Figure 9. Cognitive model of task performance for item 16: Which value of n makes the expression $\frac{5n}{2n-1}$ undefined? (NYSED, 2009a).

Figure 9 shows the TPCM for item 16. For this item, students selected a value a variable could assume in order to make a given expression undefined. Declarative knowledge played a role again, and students needed to state the conditions under which an expression is undefined. Like item 15, which depended heavily on declarative knowledge, this proved to be the most difficult item in the selected set, with seven students selecting the correct option. Once the hurdle of stating the conditions necessary for an expression to be undefined was overcome, students either set the denominator of the expression equal to zero and solved for the value of the variable, or substituted
answer options into the expression to check if the result was undefined. Like item 14, this allowed students to adopt a test-wise strategy. However, they still needed to possess the knowledge of what makes a function undefined in order to successfully respond to this item. Substitution would not result in an obviously correct option, as may have been the case in item 14. Still, this solution path differed from students who chose an algebraic approach.

Figure 10. Cognitive model of task performance for item 17: *At Genesee High School, the sophomore class has 60 fewer students than the freshman class. The junior class has 50 more than twice the students in the freshman class. The senior class is three times as large as the freshman class. If there are a total of 1,424 students at Genesee High School, how many students are in the freshman class?* (NYSED, 2009a).
The TPCM for item 17 is presented above in Figure 10. Students were asked to calculate the size of a high school freshman class based on variable expressions for the four grade levels and the total number of students in the school. Although two solution paths appear in the figure, the only difference in the approaches taken by the eight students who responded correctly to the item was whether they checked their eventual solution before matching it to an answer option. Only two students chose to check, while six students did not. It is likely that these students reasoned that the match between their calculated answer and the available options implied correctness and an additional check was not necessary. No student verbalized this thought, but students did articulate a feeling of satisfaction when their answer matched an option. This conclusion is supported by examination of the retrospective interviews, in which several students made statements such as “I saw that [my answer] was option three, so I knew that I was done.” This supports the hypothesis that student confidence was bolstered by the agreement between their work and the response option that decreased their desire or need to check their work.
Figure 11. Cognitive model of task performance for item 18: What are the vertex and axis of symmetry of the parabola $y = x^2 - 16x + 63$? (NYSED, 2009a).
The TPCM for item 18 is shown in Figure 11 (due to space constraints, Figure 11 has been split and is displayed on two pages). In this item, students were presented with the equation of a parabola in standard form and asked to identify the vertex and axis of symmetry. Nine students responded correctly to this item. One student guessed and was able to select the correct option purely by chance. Of the eight remaining students, five chose to graph the parabola using a calculator, two solved the problem with an algebraic approach, and one used both of these methods. Their specific process steps differed slightly, usually in order of steps rather than content. Still, these two distinct paths
emerged. This divergence in approaches occurred at an obvious conceptual split and the information collected about student cognitive processes was sufficient to provide robust insight into the interaction between the item and student examinees. As the actual differences in solution paths were trivial, I considered data saturation achieved despite many students adopting slightly different process paths.

This TPCM highlights the difficulty of drawing conclusions based solely on student responses to multiple-choice items. Any conclusions made about student abilities would have to be at a very coarse-grain level or would require some insight into the particular processes they undertook (e.g., requiring work to be shown). Not only did some students use the calculator while some solved algebraic expressions, but within each of these approaches, only two of eight students followed the same solution path. For example, within the algebraic approach, one student used the completing the square method to put the equation of the parabola into vertex form. Another used the axis of symmetry formula and then substituted the x-value into the original equation to calculate the vertex. Both approaches require algebraic manipulation and an understanding of quadratic equations and graphs. Still, the solution paths differ somewhat and inferences about each student based on their response would be limited.

Also of note is another instance where students were reluctant to use test-wise strategies. An inspection of the available answer options reveals that while there are only two different axes of symmetry represented across the four answer options, each option proposes a different vertex. Thus, only identifying the vertex is necessary to respond correctly to this item. Despite this, every non-guessing student correctly identified both components. For students adopting an algebraic approach, this is logical as the axis of
symmetry formula is a typical first step. However, students using a calculator had the ability to jump directly to vertex identification. This may be connected to the participants since moderately high-achieving students were selected for this study and these students may be more likely to use mathematical rather than test-wise strategies.

**Research Question 3**

The third research question concerned the comparison of expert-expected to actual student cognitive processes employed when solving the items. This question is central to the validity argument for the results of the examination. The comparison was accomplished by attempting to match the student cognitive processes recorded in the TPCMs to the process or processes described by the performance indicators within the DMCM. Cases where the process or processes evidenced in the DMCM were evident in the TPCM were considered a match.

Table 7 shows the performance indicator purportedly measured by each item and the process or processes from each model of task performance that indicated that students were solving the item in the manner expected by the developers. Put another way, when some or all of the cognitive processes shown in the TPCM match the performance indicator from the DMCM, the actual student processes are in line with expert-expected processes. For example, item 5 was intended to elicit students’ ability to “categorize data as qualitative or quantitative” (NYSED, 2009d). During the think-aloud protocols, seven students chose a solution path in which the classified each answer option as either qualitative or quantitative. Three additional students defined the term *quantitative* as involving numerical data and *qualitative* as involve data that were not numerical in nature. Then the students identified the option that offered a nonnumeric data set. In both
cases, the students demonstrated the specific skill that the item was intended to elicit. Thus, these scenarios were coded as matches between the TPCM and the DMCM.

All comparisons between the DMCM and TPCM were not as straightforward as the above example. In the task performance model for item 14, four students substituted each of the answer options into the given inequality and chose a response based on identifying the option that produces a mathematically-correct statement. The performance indicator to which this item was aligned read: “determine whether a given value is a solution to a given linear equation in one variable or linear inequality in one variable” (NYSED, 2009d). The students did display this performance. By responding correctly to the item using mathematical reasoning, they determined whether the values in the answer options were solutions to the inequality in the stem. However, it is unlikely that a teacher would instruct students in this approach. More likely, students would solve the inequality for $x$ and then evaluate the answer options. Although these students chose a path that was unlike the processes that might be required in a classroom assessment, they did perform as expected according to the aligned performance indicator.

Table 7

Item Performance Indicators and Comparable Elicited Cognitive Processes from the TPCMs (NYSED, 2009d)

<table>
<thead>
<tr>
<th>Item #</th>
<th>Performance Indicator</th>
<th>Processes from TPCM</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>A.A.27 - Understand and apply the multiplication property of zero to solve quadratic equations with integral coefficients and integral roots</td>
<td>Factor into binomials Set each equal to 0</td>
</tr>
<tr>
<td>5</td>
<td>A.S.1 - Categorize data as qualitative or quantitative</td>
<td>Classify each option as qualitative or quantitative -or- Define quantitative as numbers and qualitative as its alternative (i.e., not numbers) Identify option that does not include numbers</td>
</tr>
<tr>
<td>Page</td>
<td>A.A.45 - Determine the measure of a third side of a right triangle using the Pythagorean theorem, given the lengths of any two sides</td>
<td>Recognize application of Pythagorean Theorem Substitute leg lengths Square terms Combine like terms Take square root of both sides</td>
</tr>
<tr>
<td>------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>10</td>
<td>A.N.2 - Simplify radical terms (no variable in the radicand)</td>
<td>Break 32 into factors with 16 as a perfect square Take the square root of 16</td>
</tr>
<tr>
<td>14</td>
<td>A.A.21 - Determine whether a given value is a solution to a given linear equation in one variable or linear inequality in one variable</td>
<td>Review options for numbers less than 9 -or- Substitute answer options into inequality Eliminate options not in solution set</td>
</tr>
<tr>
<td>15</td>
<td>A.S.9 - Analyze and interpret a frequency distribution table or histogram, a cumulative frequency distribution table or histogram, or a box-and-whisker plot</td>
<td>Identify each represented data point on the plot Identify scale of plot</td>
</tr>
<tr>
<td>16</td>
<td>A.A.15 - Find values of a variable for which an algebraic fraction is undefined.</td>
<td>State conditions for expression to be undefined Set denominator equal to 0 Add 1 to both sides Divide both sides by 2 -or- State conditions for expression to be undefined Try answer options and compute value of expression Eliminate options not meeting criteria for undefined</td>
</tr>
<tr>
<td>17</td>
<td>A.A.6 - Analyze and solve verbal problems whose solution requires solving a linear equation in one variable or linear inequality in one variable.</td>
<td>Define each grade in terms of freshman class Set sum of expression = 1,424 Combine like terms Subtract 10 from both sides Divide both sides by 7</td>
</tr>
</tbody>
</table>
| 18 | A.A.41 - Determine the vertex and axis of symmetry of a parabola, given its equation (See A.G.10) | Examine table of values on calculator
Eliminate options not in table of values
-or-
Inspect graph
Identify axis of symmetry using graph
Examine table of values on calculator
Identify vertex in table of values
-or-
Identify axis of symmetry from inspection of graph
Eliminate options using axis of symmetry
Examine table of values on calculator
Identify vertex in table of values
-or-
Examine table of values on calculator
Identify vertex in table of values
Identify axis of symmetry from vertex
-or-
Use formula for axis of symmetry
Substitute for $a$ and $b$
Solve for axis of symmetry
Substitute $x=8$ into equation from stem
Solve for $y$
Identify vertex as $(8, -1)$
-or-
Use formula for axis of symmetry
Substitute for $a$ and $b$
Solve for axis of symmetry
Examine table of values at $x=8$
Identify vertex in table of values
-or-
Subtract 63 from both sides
Complete the square
Distribute the left side
Combine terms on the left side
Factor the right side
Subtract 1 from both sides
Use equation in vertex form to identify vertex
Identify axis of symmetry from x-value of vertex |

Table 8 summarizes the comparison of the performance indicators in the DMCM with the actual student processes elicited through the think-aloud protocols and reported in the item-level TCPMs. The rate of agreement between was very high. For the 75 total correct responses, 71 aligned with the DMCM for an agreement rate of 95%. Items 2, 15 and 18 showed instances of misalignment between the DMCM and the TPCM in two,
one, and one case, respectively. While the students still arrived at the correct response, they did so without eliciting the performance expected by the test developers. Overall, this small amount of mismatch demonstrates very close alignment between what experts expected students to do when presented with these test items and what the students who participated in this study actually did.

Table 8

Summary of Expert-Expected and Actual Student Processes for Algebra Items

<table>
<thead>
<tr>
<th>Item #</th>
<th>Number of Correct Responses</th>
<th>Number of Students Using Expert-Expected Processes</th>
<th>Number of Instances of Misalignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>8</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>8</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>7</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>7</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>8</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>9</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>TOTAL</td>
<td>75</td>
<td>71</td>
<td>4</td>
</tr>
</tbody>
</table>

There were two instances of misalignment for item 2. One student read the item stem, examined the answer options, concluded that she did not know how to solve the problem and selected an option by guessing. Her guess was not informed by any mathematical knowledge. Thus, it is appropriate to discount this correct response that is not aligned with expert-expected processes as a correct guess. The second instance of misalignment on item 2 is not so clear. In this case, the student recognized that a quadratic trinomial could be reduced into binomials, but chose not to factor the original equation. Instead, he substituted the answer options into binomials and distributed,
hoping to match the resulting equation with the original from the stem. He was successful in this approach. Although the student showed several different important mathematical concepts, he did not demonstrate the performance that the item was designed to elicit: understand and apply the multiplication property of zero to solve quadratic equations with integral coefficients and integral roots. Thus, his correct response was not aligned with the DMCM.

Item 15 had one instance of misalignment. According to the performance indicator, students were asked to analyze and interpret a box-and-whisker plot. The student in question simply identified the upper quartile and selected a response. The raters agreed that this process represented interpreting the plot, but not analyzing it. Since the performance indicator is conjunctive and all of the other students identified all points shown on the plot and selected the upper quartile from this list, this action was coded as representing a misalignment. This instance seems to indicate a lack of clarity in the item stem. There is no evidence that the student did anything out of the ordinary. In fact, it was the other seven students who went above and beyond the specific request of the stem. The item likely should have been written differently if the intention was to elicit both interpretation and analysis.

The final instance of misalignment was found in item 18. Similar to the first instance in item 2, a student simply guessed due to lack of knowledge about how to solve the item. There are few if any implications that can be taken away from these situations, as guessing is always a source of construct-irrelevant variance in educational assessment – especially when multiple-choice items are used (Messick, 1995). Even the high-ability students selected to participate in this study resorted to guessing. Given the number-right
scoring approach used for the examination, this is an appropriate strategy for students to adopt.

**Research Question 4**

The fourth research question asked: to what degree does the construct validity evidence provided by comparing the DMCM and the TPCMs support the claims of the exam? The claims of the Regents Examination in Integrated Algebra are not explicitly stated in the technical documentation. By piecing together statements made about the purpose of the test, the exam claim is that students who pass the exam (by scoring 65 or above) have demonstrated the minimum knowledge and understanding required by the New York State content standards to meet the exam criterion for graduation. Additionally, the claim is made that students who score at or above 85 have demonstrated mastery of these content standards (NYSED, 2009b).

Because the claims are in terms of content-area knowledge only, validation efforts should focus on the alignment between the standards that the exam is intended to measure and the actual knowledge demonstrated by students responding to the test items. The comparison of the DMCM and TPCM fits perfectly into this approach to validation. Such a comparison allows for the examination of inferences about student achievement and conceptual understanding based on the cognitive processes that students are actually using to solve item to be directly compared to the performances that experts attempted to elicit when writing the test items. To answer the fourth research question, the implications of the comparison carried out as a part of research question three must provide evidence for the claims that the test scores are intended to convey. The results of the comparison showed that out of 75 correct responses to the multiple-choice items
selected for this study, only four were arrived at through processes other than those expected by the test developers. This high rate of agreement suggests that this finding provides evidence in support of the exam claims. That is, interpreting the test results in terms of minimal competence or mastery of the content area is appropriate given the knowledge and skills elicited by the items. Still, because the stakes of this exam are so great (i.e., graduation) and the participants of this study represent only one small group of students, it was as necessary to examine the four instances of misalignment from the expected versus actual comparison when answering the third research question.

As discussed above, two of the four instances of misalignment represent possible threats to the validity of the exam claims. For item 2, the student used mathematical knowledge other than what was expected to solve the problem. Any interpretation that assumes that this student has the knowledge described by the associated performance indicator is baseless. This is not to say that the student could not have solved the item in a manner in line with what was expected by developers, it is simply evidence that it is possible to respond correctly to the item without demonstrating the aligned performance.

The instance of misalignment for item 15 is different in the mathematical sense, but shows another example of a student arriving at the correct answer without using the expert-expected processes. The student only completed only half of the expected performance and was able to correctly respond to the item. Again, the student may have been able to deliver the expected performance, but something about the item, perhaps the wording of the stem, did not elicit the desired outcome.

The fourth research question focused on the degree to which the construct validity evidence provided by comparing the DMCM and the processes in each TPCM supported
the claims of the exam. The short answer to this inquiry is: to a high degree. The 95% rate of agreement between expert-expected and actual student cognitive processes elicited by the test items supports this conclusion. Additionally, two of the four instances of misalignment are easily dismissed as construct-irrelevant variance because they came about as a result of student guessing. Only two out of 75 responses did not demonstrate exact alignment between the DMCM and TPCM.

There is a longer answer to the research question that emerges from this section. While the statistical agreement seems very high, it is important to consider not only the instances of misalignment, but the reasons for them. An item stem that does not actually elicit the desired performance is a possible culprit for item 15. Despite not being explicitly asked to, a majority of students demonstrated an analysis of the plot as well as an interpretation. The one student who did not follow this path did follow the directions of the item. Thus, it seems that clarifying the item stem is the likely rectification of this misalignment.

The misalignment in item 2 is a bit more mysterious. The student displayed mathematical knowledge, just not the knowledge expected by the item. A more holistic interpretation might conclude that the student’s approach demonstrated mathematical sophistication. However, because the exam claims are in terms of content and the associated performance indicators, this viewpoint does not contribute to the validation process. This instance of misalignment may not be easily solved by revisions to the items, but is nonetheless a threat to the validity of interpretations of the exam. Namely, the student is being credited with knowledge that he did not need to apply to respond correctly to the item. While the findings of the comparison between the DMCM and the
TPCMs do contribute significant validity evidence for the claims of the exam, it is important to make note of the instances of misalignment and use this information to improve the test development process.

Future research should investigate what interpretations can be made about solution paths that do not incorporate the expected processes (i.e., match the performance indicator), but still reflect sophisticated mathematical thinking and arrive at the correct response. Given the movement to design assessments aligned with the Common Core State Standards (CCSS, 2010), it is likely that math test items will allows for a variety of solution paths and interpretations will be made a more coarse-grained level. This presents a more holistic view of mathematics as a discipline and reflects the overarching knowledge that students must have to succeed in the years after secondary schooling, instead of focusing on individual skills that may not transfer other contexts (CCSS, 2010).
Chapter 4: Discussion

In this study, I gathered validity evidence from the results of the New York State Regents Examination in Integrated Algebra by comparing student- and expert-based cognitive models. Although comparisons of these models have occurred elsewhere in the literature, they have been limited to regression analyses investigating the predictive quality of each model. Little or no attention was given to validity evidence beyond this aspect. For this study, item-level cognitive models of students’ task performance were developed using concurrent think-aloud protocols. These models were compared with the performance indicators found in the expert-based domain mastery cognitive model used for test development.

This investigation approached validation in a manner that has not yet permeated the large-scale testing field in routine validation practices. This gap has left interpretations of student results tied to experts’ hypotheses of how students interact with the test items. As the cognitive processes that students use to solve each item can be very different than those assumed by the test developers, any interpretations made are subject to scrutiny. Through this study, I sought to provide knowledge of how students are actually operating on these test items and the cognitive processes they use which will, in turn, produce evidence for the validity or invalidity of interpretations made from the results of the examination.

Implications for Test Development

Ideally, instead of being generated concurrently with students solving the test items, the task models generated in this study would have been identified early in the test development process and used to inform all aspects of development. Indeed, Gorin
(2006) triumphs the utility of cognitive models in defining the construct of interest. However, Leighton and Gierl (2011) have repeatedly lamented that cognitive models that can inform the development of large-scale assessments are few and far between. Although cognitive models and learning progressions have been created for many content areas (e.g., Anderson, 1990; Graesser, 2007; Kilpatrick, et. al, 2001; Kuhn, 2001), the authors do not believe that they are representative of a formal testing situation because they were not designed for that purpose. Leighton and Gierl (2011) reported that “none of these models specifically reflects the types of processes that are accessed during formal testing or identifies key cognitive components used when students solve problems on achievement tests” (p. 200). For example, regardless of the generation approach employed, a cognitive model for algebraically finding the roots of a parabola would be unlikely to include the skills of substituting answer options into binomials and distributing. This is a mathematical skill that was utilized by one student, but not one typically associated with problems measuring this or similar standards. Thus it is likely that a cognitive model of task performance developed outside a testing situation may not be able to aid in the interpretation of the student’s performance. However, this viewpoint is not empirically based and even Leighton and Gierl (2011) acknowledge that non-test-based cognitive models must serve as the starting point for assessment design. Within logical limits, the content remains unchanged whether it is presented in a classroom setting or a formal testing situation.

Beyond the lack of models that account for behaviors more often seen in test taking situations (e.g., guessing, circuitous solution paths, or corrective feedback offered by a limited number of answer options (Bridgeman, 1992)), is the issue that most models
of cognition focus on high-achieving students. Put another way, cognitive models
typically represent students who have mastered a skill as opposed to those developing it.
Continued research into learning progressions has begun to account for how students
build knowledge and skills, but this has yet to widely inform models of task performance,
which still must be developed from higher-level students. A single cognitive model also
rarely covers students from different cultural backgrounds (Leighton, Cui, & Cor, 2009).
This limitation makes achieving the NRC’s mandate of a “model of how students
represent knowledge and develop competence in the subject domain” (2001, p. 2) much
trickier, as students of different achievement levels and educational backgrounds often
represent knowledge in different ways.

A major implication of the results of this study is the reinforcement of the utility
of expert judgment when designing test items to measure performance indicators. For the
selected items, the majority of student solution paths employed cognitive processes
indistinguishable from those expected by the experts involved with item development.
This agreement provided evidence that expert judgment was a trustworthy approach to
ensuring that these items elicited the desired performance. In New York State, all test
items are written by certified teachers who undergo a training session. Perhaps the
familiarity that these teachers have with student problem-solving behaviors and the
content standards provide an ideal position for ensuring proper alignment between the
two. In a different development process, perhaps where items were written by content
experts who were not state-certified teachers, there might be more instances of
misalignment between the DMCM and TPCMs. The approach undertaken by NYSED
was effective in ensuring that items measure the intended performance for high-achieving students for the selected items in this study.

**Implications for item writing.** Although New York has a development process that seems to produce positive results for the selected items where expected and actual processes are concerned, cognitive models still have the potential to inform, and perhaps improve, item writing. Beyond her discussion on how models of cognition can aid construct definition, Gorin (2006) suggested that assessments designed around a cognitive framework (i.e., using one or more cognitive models as the basis for construct definition) introduce the possibility for innovative item types which support diagnostic inferences and the possibility of automating item generation. Of course these would require the development and validation of cognitive models specifically for the intended purposes. The TPCMs developed as a part of this investigation are limited in the diagnoses they can provide because they deal only with correct responses to the items.

Still, the implications of the TPCMs for item development are more than a verification of the alignment abilities of NYSED’s item writers. The models, if generalized to the performance indicator, can be used to ensure future items written to elicit the specific performance take advantage of knowledge of common student solution paths. Additionally, the instances of misalignment between the DMCM and TPCMs can serve as formative feedback and help the writers improve their already admirable calibration to student behaviors. Yet, it is not only the results of this investigation that have the potential to inform item development of future Regent Examinations, but the approach as a whole. One area where elicitation of student cognitive processes can contribute to test development is item field testing.
Like most producers of large-scale achievement tests, NYSED field tests all items before they appear on an operational exam. This provides the opportunity to examine classical and item response theory statistics, as well as raw student work for constructed response items. Determinations about item quality can be made using this information and items can be thrown out, revised and re-field tested, or accepted for operational testing based on statistical and other criteria. The think aloud protocols and subsequent coding into TPCMs could be used as an early stage of field testing. Under current NYSED processes, items are written, revised, and then sent out statewide on standalone field test forms to a representative sample of approximately 1,000 examinees per item (NYSED, 2009b). The methods from this investigation could be inserted after revision and before full-scale field testing.

Before expending the resources necessary to collect response data from hundreds if not thousands of students, think-aloud protocols could be conducted on a much smaller sample. Almond et al. (2009) suggested anywhere from five to twenty students may be appropriate as long as data saturation is achieved. Although this would be a time consuming endeavor, the resulting information about student interaction with the test items would be invaluable. Deficiencies in the alignment between the performance elicited by the items and the indicator they are written to would become evident very quickly. Additionally, sources of construct irrelevant variance such as clueing in the item stem, problems with the distractors, or unintended corrective feedback could be identified directly instead of inferred from statistical results. The data received from the larger field testing is valuable – especially as NYSED uses these statistics for operational form equating. However, the additional effort to conduct think-aloud protocols to investigate
item quality would be well worth it for test developers at NYSED or any other large-scale test construction company. Demonstrating the value of such an approach, Kaliski and colleagues (2011) described a similar approach to informing item development undertaken by The College Board for its AP Exams.

Implications for Validation

The earlier discussion surrounding the fourth research question looked specifically at what the results of this investigation could contribute to the argument for or against the validity of interpretations made from the Regents Exam in Integrated Algebra. The results of this exam are interpreted in terms of student achievement in the construct of Integrated Algebra as defined by NYSED content standards. Students are classified into performance levels that reflect their performance in the content area (e.g., Proficient, Mastery, etc.). The major claim of the exam that requires evidence for validity is that students who pass the exam have demonstrated the minimum knowledge and understanding required by the New York State content standards for graduation and students who score at or above 85 have demonstrated mastery of these content standards. This study was conducted with the intention of seeking validity evidence for the Integrated Algebra Regents Exam, but the implications for future validation studies extend beyond the specific test selected for this investigation.

As with the implications for test and item development, it is the overall process undertaken in this study that can help inform future validation work. Kane’s (2006) argument-based approach to validation provides an excellent opportunity for cognitive process verification to be used as evidence supporting the interpretation/use argument. Due to the NRC call (2001) for models of student cognition to underlie achievement
tests, the IUA and claims for the exam are ideally evident in whatever model is used for development. Even in instances where an established and empirically-validated model of cognition informs test development and results interpretation, the reality exists that it is not known exactly how students are interacting with the items in a testing situation. This lack of knowledge opens the door to conducting think-aloud protocols with students in a test-taking scenario to better understand this interaction and its implications for interpreting the results of the exam.

Once the IUA is established, it becomes the responsibility of developers to offer evidence in support of the proposed interpretations and uses as part of the validity argument. This is where an examination of actual student cognitive processes can offer such evidence. For example, if developers expect to make decisions about a student’s understanding of algebra and think-aloud protocols reveal that all test items can be solved without foundational algebraic concepts, the interpretations from the exam would be called into question. At the item level, instances of misalignment between standards or performance indicators (i.e., claims of a finer grain-size) and actual student work would suggest that the item was not operating as expected. In this case, the interpretations from that specific item would not contribute to the overall interpretations to be made from the full test, which would again threaten the validity of the IUA.

As detailed in the third chapter, NYSED relies heavily upon statistical indicators as evidence for reliability and validity. Additionally, the involvement of certified teachers in all aspects of test development is intended to establish evidence that the test accurately measures the intended content. This approach is more in line with perspectives on validity that are not unified in the viewpoint of Messick (1989) or Kane (2006, 2013).
Cronbach (1960) proposed that content validity, one of four types of validation, referred to asking the question: “Does this test represent the content or activities I am trying to measure?” (p. 104). Then the test items would be compared to the desired content in an effort to answer this question. Thorndike and Thorndike-Christ (2010) described expert validity as a subset of content validity in which the judgment of content experts was used to answer questions surrounding the alignment between the construct of measurement and the test items. NYSED (2009c) invokes the sentiment of content validity as ensuring that the Regents Examination “adequately samples the relevant material it purports to cover” (p. 14).

As the field has moved beyond the fractured view of validity (see Messick, 1995), it is crucial that NYSED and other large-scale test developers keep up with current advances in validity theory and inference validation. Kane’s argument-based approach represents the current best practice in validation and NYSED would be wise to adopt such a course in order to fully defend the exam results. As previously mentioned the exam claims are not entirely clear. The first step in adopting a strong course of validation would be to explicitly state these claims. Without clarity, it is difficult to collect supporting evidence. In Kane’s (2013) terms, the IUA must be “coherent and complete (in the sense that it fully represents the proposed interpretation or use)” (p. 2).

Once exam claims are established as a part of the IUA, they must be supported with evidence. Kane pointed out that the robustness of this evidence can vary in relation to the ambitiousness of the claim. The claim of the Regents Examination in Integrated Algebra is one of domain mastery over a set of content standards. As this construct is not readily observable, appropriate evidence for the IUA must move beyond students’
abilities to perform the tasks included on the exam. The only current, non-expert-based evidence that NYSED (2009c) provides is a principal component analysis that shows that the main factor accounts for 80% of the variance among factors. Using this result, the developers conclude that “the test has 1 dominant underlying construct” (p. 22). Only expert judgment is available to suggest that this dominant construct actually is the construct of interest (i.e., algebra knowledge). Although the domain mastery claim is made at the test level (as opposed to the item level), there is an additional need to offer evidence that the test items are accessing the construct of interest. As described above, misalignment between the target of measurement for an item and actual item performance threatens the fidelity of the IUA which is an argument for interpreting the results of the full test, not specific items. Thus, ensuring appropriate inferences are made at the item level is crucial for gathering evidence to support interpretations of the test.

Because validation is a process that begins with design and development, introducing think-aloud protocols into the course of validation would provide additional evidence that the students are operating on the test items in such a way as to demonstrate skills in the construct of interest. Additionally, such an approach would help ensure that it is algebra that is the major factor in student performance. Kane (2013) noted that supporting every claim in the IUA with empirical evidence is not a realistic goal. Because think-aloud protocols can contribute to the validity argument in several ways, this approach seems efficient by providing the greatest reward for the resources invested.

Implications for Score Reporting

Score reporting was not a core focus of this study, but there are several implications for reporting results. As Gierl, Leighton, and Hunka (2007) wrote, “test
score reporting is the primary method of providing information about the meaning and possible interpretations of test results to all stakeholders” (p. 264). Goodman and Hambleton (2004) published possibly the most comprehensive review of score reporting practices available. In their report, they used four questions established by the National Education Goals Panel (NEGP; 1998). Two of these questions are relevant to the information that can be drawn, and thus reported, from the results of a test developed from TPCMs such as those developed in this study: “What types of skills or knowledge does [my child’s] performance reflect?” and “What can I do to help my child improve?” (p. 36). Without clear articulation of the knowledge and skills elicited by test items and evidence that the items are operating as intended, it would not be possible to provide such diagnostic information. However, when models of cognition have been developed, validated, and used to inform item and test development, it becomes possible to report student results in terms of specific knowledge and skills and to provide information on what students can do to improve in the content.

Huff and Goodman (2007) elaborated on the demand for diagnostic information. Teachers, parents, and students all desire a clearer picture of what their test results mean in terms of knowledge and skills. Unfortunately, as the authors stated, “test results are not necessarily connected to classroom learning and instruction, and accordingly, have limited utility for educators and students” (p. 22). Specifically, teachers hope to use the test result data to inform their instructional practices and provide relevant supports to students. To the degree that test results reflect knowledge and skills, improving student achievement in these areas will result in higher test scores. Although this overly-
simplified view is impacted by multiple sources of construct-irrelevant variance, the sentiment is the driving force behind many educational reform movements.

Although the Regents Examination in Integrated Algebra is not intended to be a cognitively diagnostic assessment (see Leighton & Gierl, 2007b), underlying the exam with principles from the learning sciences will allow for finer-grain inferences in terms of specific student knowledge and skills. Before the TPCMs were created and retrofit in this study, interpretations were limited to dichotomous (yes/no) designations for each performance indicator. Such interpretations are likely to be unreliable as they are based on only one item (Goodman & Hambleton, 2004). Still, at an aggregate level (e.g., all students with the same teacher), there may be some limited information available about student knowledge and skills. With TPCMs like those in this study, items could be developed at a finer grain size and the information available to teachers more robust. For example, an item could be written to elicit a specific solution path or multiple items could be grouped to form a subdomain. Task models at the test or subdomain level would be of great help to stakeholders in making interpretations about specific knowledge or skills in the construct.

**Limitations**

The limitations of the study include small sample size for both students and items. The full-length operational form of the Regents Examination in Integrated Algebra includes 37 items and is taken by 300,000 students annually. Precise conclusions about the full exam or the examinee population cannot be drawn. The generalizability of the results to a larger population is limited due to the participants selected for the investigation. Only moderately high-achieving students were included in this study as the
relevant literature deemed it appropriate to exclude students who might not be able to coherently verbalize their mathematical cognitive processes. Thus the resulting TPCMs may only apply to these higher-level students. However, these models can serve as a starting point for future validation studies including a wider range of students. They may also be used to guide integrated algebra item development.

Generalizability of results may also be limited by the items selected. The analyses were conducted at the item level, but the implications for validation apply to the full test. Although it is true that issues with even one item impacts the validity of test interpretations, a full course of validation would cover all items or at least a larger, representative sample. The items selected for this investigation were chosen based on statistical criteria rather than content representation. The selected items may not be a good sample of the range of content assessed by the exam. Additionally, only multiple-choice items were analyzed in this study. Further investigation using this type of approach to validation would need to examine student responses to constructed-response items as well.

Although the methods for coding followed standard practices employed in qualitative research including calibration across two raters, the interpretations of student processes are those of the raters. Both raters had extensive experience with STEM course work, but it was necessary at times for them to interpret student processes and expert-intentions based on their best judgment. Again, methodological precautions were followed to ensure data trustworthiness, but, like any qualitative investigation, the results must be interpreted in terms of the judgment of experienced content experts and researchers. This is similar to how item writers must rely on their expertise and
knowledge of student problem-solving behavior when designing a test item to elicit a specific performance.

The process examinees must follow when thinking aloud presents a limitation. The best practices for think-aloud protocols were followed rather than the realistic constraints of the testing environment. Participants were given unlimited time to solve items whereas a testing situation would restrict the amount of time available. During an operational test, students might draw early conclusions or rely on guessing if time were to run short. The fact that the testing situation is not authentic may reduce veridicality, that is, accuracy of self-reports of cognitive processes (Russo, Johnson, & Stephens, 1989). However, Leighton (2004) suggested that concurrent reports do offer some protection against this problem.

Finally, there are two alternative hypotheses for the high degree of agreement between the DMCM and the TPCMs that should be considered. One is the possibility that the performance indicators are fine-grain enough to be prescriptive in terms of what students must do to solve items purporting to measure these skills. In other words, the only way to respond correctly to an item measuring a certain performance indicator is in the manner dictated by the indicator. For example, the performance indicator for item 10 is “simplify radical numbers.” It is difficult to think of a way in which a student could demonstrate this skill without being asked to simplify a radical number and then successfully completing the task.

A second hypothesis is that the test items are not accurately measuring student achievement in terms of the performance indicators due to flaws or limitations in the items themselves. There is no evidence that this was true for the items selected for this
study. Although guessing was not widespread in this investigation, there were two instances where students guessed the correct option. In item 2, a student identified a number in the options that was similar to the equation in the stem. The student based their guess on this similarity and was rewarded with a correct answer. It is not clear if this should be classified as truly an item flaw or an unfortunate coincidence, but it is obvious that the item did not measure the intended skill for this student.

**Future Research Directions**

Much of the research that informed this study is from the realm of cognitive diagnostic assessment (e.g., Leighton & Gierl, 2007b; Nichols, 1994). Although the Regents Examination in Integrated Algebra is a summative achievement test, there are multiple principles from cognitive diagnostic assessment that can help improve the design of the exam as well as interpretations made from the results. Leighton and Gierl (2007b) describe cognitive diagnostic assessments as those that are “designed to measure the specific knowledge structure (e.g., distributive rule in mathematics) and processing skills (e.g., applying the distributive rule in appropriate mathematical contexts) an examinee has acquired” (p. 146). This is accomplished by basing test development on empirically validated cognitive models of task performance, rather than on content and assessment experts using established psychometric frameworks.

Given the effort needed to develop and validate task performance models, it is easy to see why large-scale testing programs are not currently using this approach. However, this may be an instance in which the initial effort will be rewarded in the payout. Current large-scale tests allow for limited inferences. Typically, this is considered sufficient because large-scale exams are tasked with establishing more coarse-grain
determinations (e.g., “pass”). Still, as test results are called upon to deliver more and more information (Warner, 2013), it seems worth the work to develop tests that are based on psychological frameworks and can report on “students’ misconceptions and learning in general” (Leighton & Gierl, 2007b, p. 150). Such assessments span the gap between summative and formative designations by allowing for inferences on student knowledge and skills that can be used to make the types of decisions that need to be made at the end of a course, but also providing information about individual students’ cognitive strengths and weaknesses.

Certainly additional research should be conducted on the approaches outlined in this investigation and discussed above. Think-aloud protocols hold promise for informing item development as well as validation of inferences and should be used as a primary methodology in delving more deeply into how actual student cognition can aid in these processes. Score reports can be improved to better inform education stakeholders what a student’s performance actually means. However, all of this is contingent upon a new breed of tests. These new assessments must be based on empirically validated models of student learning and cognition. Those in the field of learning science are rapidly expanding the available resources and advancing knowledge about how students think and learn.

As the United States embarks on a national reform agenda that seeks to standardize expectations in the hope that all students can be on track to college and careers (CCSSI, 2010), consortia are busy developing assessments that intend to measure students’ preparation on this trajectory (Smarter Balanced Assessment Consortium (SBAC), 2013; Partnership for Assessment of Readiness for College and Careers...
With the goal of measuring student progress, tests developed by these consortia must not only provide information on students’ current level of knowledge and skills (i.e., summative), but also give direction to teachers and schools for improving instruction (i.e., formative). Adopting a cognitive assessment framework for test development may be the ideal approach to ensuring that both outcomes are possible and productive for stakeholders. Indeed PARCC has developed evidence statements to specifically detail the knowledge and skills that items will elicit. Although these statements will require refinement and validation, likely with think-aloud protocols in a manner similar to this study, this is a promising beginning to designing a test that can meet the challenges laid out over the past 25 years by Snow and Lohman (1989), the NRC (2001), and Leighton and Gierl (2011). In their latest call for incorporating cognitive models into assessment design Leighton and Gierl referred to the joining of the learning sciences and educational measurement as a marriage. This metaphor seems appropriate to the monumental task ahead of the field. Like any happy marriage, it is expected that there will be bumps along the road and wrong turns taken, but the result will be well worth it in the end.
References


Appendix A: Interview Protocol

Subject ID: ___________

Interview Protocol

1. Collect and review parental consent documents (see Appendix B)
2. Student assent procedure and documentation (see Appendix B)
3. Pre-interview survey
4. Read interview script
5. Warm-up items
6. Cognitive process interview
7. Retrospective Interview
8. Solicitation of feedback
9. Conclusion

Collect/Review Parental Consent Documents

Students who do not have parental consent are not eligible for participation in this study.

Student Assessment Procedure and Documentation

Students will be provided with assent documentation and the researcher will address any questions or concerns. Students who agree to take part in the study will sign the assent documents and the interview can begin. The video recorder will be turned on at this point.

Interview Script

Read aloud to student:

Today is (date) and this is an interview of subject (ID number) conducted by (researcher). It is (time) and we are conducting this interview (location).
Thank you for taking part in this study. Please let me remind you that participation is voluntary and you are free to discontinue at any time. Also, you are free to ask questions or raise concerns at any point in your participation. Today’s study should take no more than one and one-half hours.

Today you will be asked to solve some multiple-choice Integrated Algebra problems. I know that you have recently taken the New York State Regents exam in Integrated Algebra. These items are from a previous exam, not the test that you took. While you should try to treat this like a testing situation, you are not being graded and I am actually just as interested in how you get to your final answer as I am the response that you select.

For this study, I would like you verbalize, or say out loud, every thought that you have while reading the item, engaging in problem solving, and selecting a response. I will emphasize that you should tell me every thought that you have from the time you first see the item to when you are finished. We will do this for nine multiple-choice items. Remember, there are no right or wrong thoughts – you should just say what you’re thinking in terms of ideas, problem solving strategies, etcetera.

Please talk aloud constantly. Do not try to plan what you would like to say or worry that what you are thinking may not be worth saying if it does not help solve the problem. You also do not need to reflect on what you are saying or try to explain it to me. You can pretend that I am not here and you are talking to yourself or thinking aloud. The important thing is that you keep talking. If time passes and you don’t say anything, I will remind you to talk about what you are thinking. However, you can take as much time as you want to solve each item.

Do you understand what I am asking you to do?

Everything we do today will be recorded so that there is an accurate record of your verbalized thoughts. Anything you say will be kept confidential and anonymous. Do you have any questions or concerns?

Warm-up Items

*Read aloud to student:*

I am about to model a think-aloud procedure for you so you can see what I’m asking you to do. I will solve a sample algebra problem while verbalizing the thought processes that I am using to respond. Please observe how I do this so that you understand what the process involves. Do you have any questions before I begin? I will ask you if you have any questions after I finish as well.

- Demonstration by researcher
Read aloud to student:

Do you have any questions about what you just observed? (Answer any questions they raise.)

Now I am going to ask you to verbalize your thinking, or thinking-aloud, while you solve a sample algebra item. Use the model that I just provided as a guide. If you are not sure if you should say something out loud, remember that I am interested in all of your thoughts, so you should go ahead and say whatever you are thinking. I will give you some feedback after you finish so that we can go through the other items as effectively as possible. Do you have any questions?

- Sample item for student

(Give student feedback based on think-aloud frequency/quality/clarity)

Read aloud to student:

Now that you have had a chance to see a model think-aloud, try thinking-aloud for a sample item, and receive feedback on your verbalizations of thoughts, we are ready to move onto the algebra items for the study. Do you have any questions before we begin?

Cognitive Process Interview

Student will work through nine multiple-choice algebra items and verbalize all of his or her thought processes while responding to the items. Researcher will observe the student and take notes on verbalized processes. Researcher will not interact with the student except to answer any questions unrelated to content or thought processes, help with procedural issues (e.g., locating next item, calculator, or other material), or prompt students if silent for more than 30 seconds.

Sample Prompts:

“What are you thinking?”
“Please think aloud.”
“I am interested in what you’re thinking right now.”

Retrospective Interview

Read aloud to student:

Thank you for your work on those items. What I would like to do now is have you take another look at the work you have done. Let’s go through each item again, but since you have already solved them, you do not need to do that again. What I would like you to do this time is remember what your thought processes were when you were solving that item. You can say things like: “When I first saw the question, I thought it was going to
require multiplication because I saw the word ‘product’ and I know that that is the answer to a multiplication problem.”

Take as much time as you would like to reflect on each step of each item and think or remember out loud what thought processes you were having back when you solved the item the first time. I will not interrupt you during this time and all I ask is that you do your best to remember what you were thinking. The reason that we are going through this again is to see if there were any thought processes that you had but forgot to say out loud. Do you have any questions about what I am asking you to do?

**Solicitation of Feedback**

*Read aloud to student:*

Thank you for taking part in this study. We are done with the data collection process and I have no more questions for you. Do you have any questions for me or anything else that you would like to share?

**Conclusion**

Once the student has shared any feedback and had any final questions answered by the researcher, the researcher will conclude the interview and stop the video recorder.
Appendix B: Consent Documentation

Zachary B. Warner
University at Albany
Albany, NY 12222
Phone: (518) 256-2165

Dear Parents:

My name is Zach Warner and I am a Ph. D. student in the Educational Psychology program at the University at Albany. I am seeking your permission for your child to participate in an investigation into the cognitive processes underlying test items on the NYS Regents Examination in Integrated Algebra. To be eligible for participation in this study, your child must have taken last year or currently be taking an algebra course that leads up to the NYS Regents Examination in Integrated Algebra.

At an interview time and location that we arrange, I will ask your child to solve nine-multiple-choice items taken from the June 2009 Integrated Algebra Exam while describing their thought processes verbally. This process will be video recorded and will likely take between 60 and 90 minutes. I hope to discover the cognitive processes that they use to solve these test items and check the alignment between NYS expert-anticipated process and actual student-reported processes. There are no anticipated risks associated with involvement in this investigation beyond your child feeling nervous about trying something new (thinking-aloud). For their participation, you child will received a $25 gift card redeemable at Amazon.com. This study has the formal approval of the Institutional Review Board of the University at Albany.

Findings of the study will be available upon request and will be published as my doctoral dissertation. However, no confidential information will be collected including your child’s name. All information obtained in this study is strictly confidential unless disclosure is required by law. In addition, the Institutional Review Board and University or government officials responsible for monitoring this study may inspect these records. Once the data are collected, it will benefit others by helping investigate the validity of the Regents Examination in Integrated Algebra.

Participation in this project is voluntary. Even after you agree to your child’s participation in the research or sign the informed consent document, they may decide to leave the study at any time. All youth whose parents have consented to their child’s participation will be given a consent form that they must sign to participate in the study. Those that choose not to participate may hand the form in blank and will not be asked to take the survey. Even those who
consent may skip any questions they choose or may discontinue the survey at any time.

Your Rights as a Participant
If you have any questions concerning your child’s participation in this research or if you wish to report any concerns about the study, please contact the University at Albany’s Office of Regulatory Research Compliance at its toll-free number 1-866-857-5459 or via email at hsconcerns@albany.edu.

Please fill out the form at the bottom of this letter and return it to me indicating your choice. You may also return the form blank in which case your child will not be asked to participate. One copy of this document will be kept together with the research records of this study. Also, you will be given a copy to keep. If you have any questions about the study, I can be reached by phone at (518) 256-2165 or by email at zwarn@albany.edu. My faculty advisor is Dr. Heidi Andrade who may be reached at (518) 442-5055.

Yours truly,

Zachary B. Warner, M.S., M.S.T.
Graduate Student
Educational Psychology Program

If you consent to your child’s participation, please print their name below and sign where indicated. You may also return this form blank. In that case, your child will not be contacted to take part in the study.

I have read, or been informed of, the information about this study. I hereby consent to my child’s participation in the study. (full name)__________________.

__________________________
(Parent Signature)

DO NOT SIGN THIS FORM IF IT IS NOT STAMPED BY THE INSTITUTIONAL REVIEW BOARD, OR IF THE EXPIRATION DATE HAS PASSED.
Appendix C: Student Assent Form

**STUDENT ASSENT FORM**

<table>
<thead>
<tr>
<th><strong>Title of the study</strong></th>
<th>Comparing of Expert- and Student-Based Cognitive Models in Integrated Algebra.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Purpose of the study</strong></td>
<td>To help me understand how you approach problem-solving on the Algebra Regents and see if your way is the same as the math experts that work for New York State.</td>
</tr>
<tr>
<td><strong>Eligibility</strong></td>
<td>To be a part of this study, you must have taken last year or currently be taking a course that leads up to the NYS Integrated Algebra Regents Exam.</td>
</tr>
<tr>
<td><strong>Participation</strong></td>
<td>Your parents have given their permission for you to participate, but the ultimate choice about whether to participate or not is yours.</td>
</tr>
<tr>
<td><strong>What you will do</strong></td>
<td>We will set up and interview time and location that is convenient for you and your parents. I will ask you to think-aloud while solving nine multiple-choice math problems so that I can observe the steps you take to get to the answer. This interview will be videotaped.</td>
</tr>
<tr>
<td><strong>Incentive</strong></td>
<td>For your participation, you will receive a $25 Amazon.com gift card.</td>
</tr>
<tr>
<td><strong>Withdrawal</strong></td>
<td>You do not have to participate in this study. You can participating at any time. If you want to stop participating, just tell me.</td>
</tr>
<tr>
<td><strong>Grades</strong></td>
<td>The Regents exam is from June 2009 and will not count toward a grade. Participating in the study, not participating, or quitting the study will have no effect on your grades.</td>
</tr>
</tbody>
</table>
**Risks and Benefits**

There are few risks of participating. You might feel uncomfortable trying new things, such as thinking-aloud while solving your math problems, but I will demonstrate how to do it and you can ask questions at any time. The benefit of participating in this study is learning about the processes you use to solve math problems.

---

**I volunteer to be a part of this study.**

Your ID number (assigned for this project only): ___________ Today’s Date: ___________

---

**Investigator contact information:**

Zachary B. Warner, M.S., M.S.T.  
University at Albany  
(518) 256-2165  
zwarner@albany.edu

---

**Faculty advisor contact information:**

Heidi Andrade, Ed. D.  
University at Albany  
(518) 442-5055  
handrade@albany.edu

---

One copy of this document will be kept together with the research records of this study. Also, you will be given a copy to keep. All information obtained in this study is strictly confidential unless disclosure is required by law. In addition, the Institutional Review Board and University responsible for monitoring this study may inspect these records.

---

**Your Rights as a Participant**

If you have any questions concerning your participation in this research or if you wish to report any concerns about the study, please contact the University at Albany’s Office of Regulatory Research Compliance at its toll-free number 1-866-857-5459 or via email at hsconcerns@albany.edu.
Appendix D: Script for Recruitment of Student Participants

Thank you for agreeing to meet with me. My name is Zach Warner and I am a doctoral student in educational psychology at the University at Albany. I am working on a research project called *Comparing of Expert- and Student-Based Cognitive Models in Integrated Algebra*. The purpose of this study is to help me understand how you and other students approach problem-solving on the Integrated Algebra Regents and see if your approaches to math problems are the same as the math experts that work on the Regents exams for New York State.

I have asked you to consider participating because you either took an algebra course in school last year or are taking one now that leads up to the NYS Integrated Algebra Regents Exam. If you decide to take part in this study, we will set up and interview time and location that is convenient for you and your parents. Then, I will ask you to think-aloud while solving nine multiple-choice math problems so that I can observe the steps you take to get to the answer. This interview will be videotaped.

Your parents have signed a form saying that it is alright with them for you to participate in this study if you want to, but the decision is up to you. You do not have to participate and you can also start participating and then decide to stop; all you have to do is tell me if you want to stop. As compensation for your time and work today, you will receive a $25 Amazon.com gift card.

There little risk associated with participating in this study. You might feel uncomfortable trying new things, such as thinking-aloud while solving your math problems, but I will demonstrate how to do it and you can ask questions at any time. The main benefit of participating in this study is learning about the processes you personally use to solve math problems. It is called *metacognition* when you think about your own thinking and being metacognitive can help you succeed in school and in other activities.

Do you have any questions or concerns about the study?

Would you like to participate in this research?
Appendix E: Integrated Algebra Items Selected for Study

Note: All items are taken from the June 2009 NYS Regents Examination in Integrated Algebra.

2 What are the roots of the equation $x^2 - 7x + 6 = 0$?
   (1) 1 and 7  (3) −1 and −6
   (2) −1 and 7  (4) 1 and 6

5 Which data set describes a situation that could be classified as qualitative?
   (1) the ages of the students in Ms. Marshall’s Spanish class
   (2) the test scores of the students in Ms. Fitzgerald’s class
   (3) the favorite ice cream flavor of each of Mr. Hayden’s students
   (4) the heights of the players on the East High School basketball team

9 What is the value of $x$, in inches, in the right triangle below?

![Right triangle diagram]

   (1) $\sqrt{15}$  (3) $\sqrt{34}$
   (2) 8  (4) 4
10 What is $\sqrt{32}$ expressed in simplest radical form?

(1) 16\sqrt{2} \hspace{1cm} (3) 4\sqrt{8}
(2) 4\sqrt{2} \hspace{1cm} (4) 2\sqrt{8}

14 Which value of $x$ is in the solution set of $\frac{4}{3}x + 5 < 17$?

(1) 8 \hspace{1cm} (3) 12
(2) 9 \hspace{1cm} (4) 16

15 The box-and-whisker plot below represents students’ scores on a recent English test.

```
                |
                |
                |
                |
                |
                60 70 80 90 100
```

Student Scores

What is the value of the upper quartile?

(1) 68 \hspace{1cm} (3) 84
(2) 76 \hspace{1cm} (4) 94

16 Which value of $n$ makes the expression $\frac{5n}{2n - 1}$ undefined?

(1) 1 \hspace{1cm} (3) $-\frac{1}{2}$
(2) 0 \hspace{1cm} (4) $\frac{1}{2}$
17 At Genesee High School, the sophomore class has 60 more students than the freshman class. The junior class has 50 fewer students than twice the students in the freshman class. The senior class is three times as large as the freshman class. If there are a total of 1,424 students at Genesee High School, how many students are in the freshman class?

(1) 202  (3) 235
(2) 205  (4) 236

18 What are the vertex and axis of symmetry of the parabola $y = x^2 - 16x + 63$?

(1) vertex: $(8, -1)$; axis of symmetry: $x = 8$
(2) vertex: $(8, 1)$; axis of symmetry: $x = 8$
(3) vertex: $(-8, -1)$; axis of symmetry: $x = -8$
(4) vertex: $(-8, 1)$; axis of symmetry: $x = -8$
### Appendix F: Item Alignment Map Obtained from NYSED

June 2009 Regents Examination in Integrated Algebra Test Map (NYSED, 2009d)

<table>
<thead>
<tr>
<th>Item Position</th>
<th>Item Type</th>
<th>Maximum Credits</th>
<th>Content Strand</th>
<th>Performance Indicator</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Measurement</td>
<td>A.M.1</td>
</tr>
<tr>
<td>2</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Algebra</td>
<td>A.A.27</td>
</tr>
<tr>
<td>3</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Algebra</td>
<td>A.A.12</td>
</tr>
<tr>
<td>4</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Algebra</td>
<td>A.A.1</td>
</tr>
<tr>
<td>5</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Statistics and Probability</td>
<td>A.S.1</td>
</tr>
<tr>
<td>6</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Algebra</td>
<td>A.A.4</td>
</tr>
<tr>
<td>7</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Algebra</td>
<td>A.A.25</td>
</tr>
<tr>
<td>8</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Statistics and Probability</td>
<td>A.S.21</td>
</tr>
<tr>
<td>9</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Algebra</td>
<td>A.A.45</td>
</tr>
<tr>
<td>10</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Number Sense and Operations</td>
<td>A.N.2</td>
</tr>
<tr>
<td>11</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Measurement</td>
<td>A.M.2</td>
</tr>
<tr>
<td>12</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Algebra</td>
<td>A.A.7</td>
</tr>
<tr>
<td>13</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Algebra</td>
<td>A.A.23</td>
</tr>
<tr>
<td>14</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Algebra</td>
<td>A.A.21</td>
</tr>
<tr>
<td>15</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Statistics and Probability</td>
<td>A.S.9</td>
</tr>
<tr>
<td>16</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Algebra</td>
<td>A.A.15</td>
</tr>
<tr>
<td>17</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Algebra</td>
<td>A.A.6</td>
</tr>
<tr>
<td>18</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Algebra</td>
<td>A.A.41</td>
</tr>
<tr>
<td>19</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Geometry</td>
<td>A.G.3</td>
</tr>
<tr>
<td>20</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Geometry</td>
<td>A.G.6</td>
</tr>
<tr>
<td>21</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Algebra</td>
<td>A.A.16</td>
</tr>
<tr>
<td>22</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Algebra</td>
<td>A.A.34</td>
</tr>
<tr>
<td>23</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Algebra</td>
<td>A.A.13</td>
</tr>
<tr>
<td>24</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Geometry</td>
<td>A.G.8</td>
</tr>
<tr>
<td>25</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Algebra</td>
<td>A.A.10</td>
</tr>
<tr>
<td>26</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Number Sense and Operations</td>
<td>A.N.1</td>
</tr>
<tr>
<td>27</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Number Sense and Operations</td>
<td>A.N.4</td>
</tr>
<tr>
<td>28</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Measurement</td>
<td>A.M.3</td>
</tr>
<tr>
<td>29</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Algebra</td>
<td>A.A.17</td>
</tr>
<tr>
<td>30</td>
<td>Multiple-Choice</td>
<td>2</td>
<td>Algebra</td>
<td>A.A.29</td>
</tr>
<tr>
<td>31</td>
<td>Constructed-Response</td>
<td>2</td>
<td>Number Sense and Operations</td>
<td>A.N.8</td>
</tr>
<tr>
<td>32</td>
<td>Constructed-Response</td>
<td>2</td>
<td>Algebra</td>
<td>A.A.20</td>
</tr>
<tr>
<td>33</td>
<td>Constructed-Response</td>
<td>2</td>
<td>Statistics and Probability</td>
<td>A.S.18</td>
</tr>
<tr>
<td>Item Position</td>
<td>Item Type</td>
<td>Maximum Credits</td>
<td>Content Strand</td>
<td>Performance Indicator</td>
</tr>
<tr>
<td>---------------</td>
<td>----------------</td>
<td>-----------------</td>
<td>---------------------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>34</td>
<td>Constructed-Response</td>
<td>3</td>
<td>Geometry</td>
<td>A.G.1</td>
</tr>
<tr>
<td>35</td>
<td>Constructed-Response</td>
<td>3</td>
<td>Algebra</td>
<td>A.A.9</td>
</tr>
<tr>
<td>36</td>
<td>Constructed-Response</td>
<td>3</td>
<td>Statistics and Probability</td>
<td>A.S.8</td>
</tr>
<tr>
<td>37</td>
<td>Constructed-Response</td>
<td>4</td>
<td>Algebra</td>
<td>A.A.44</td>
</tr>
<tr>
<td>38</td>
<td>Constructed-Response</td>
<td>4</td>
<td>Statistics and Probability</td>
<td>A.S.5</td>
</tr>
<tr>
<td>39</td>
<td>Constructed-Response</td>
<td>4</td>
<td>Geometry</td>
<td>A.G.9</td>
</tr>
</tbody>
</table>
Appendix G: Coding Strategy, Policy, and Process for Calibration

Coding Strategy

Begin coding after the student reads the item. Code each student’s cognitive processes using the audio file, transcript, and original student work as sources. The coding should take place at the level of individual cognitive processes. For example, a student solving $2x + 1 = 7$ for $x$ might use the following steps:

1. Subtract 1 from both sides of the equation
2. Divide both sides of the equation by 2
3. Select option $x = 3$

The overarching work may seem to be finding the solution to an algebraic equation, but the cognitive processes within are the desired level of coding.

Any additional processes elicited during the students’ retrospective reports will be included in the solution path because the processes were used while responding to the item.

Coding Policy

- The process must be explicit in the students’ words or writing.
- If necessary, use context (i.e., the type of mathematics being performed) to provide clues for the specific mathematical cognitive process – especially if the student struggles with vocabulary, but performs a mathematical step on their paper.
- If the student attempts to solve the item one way but then changes direction, code all steps.
- Ignore “I don’t know” or similar phrases as some students may say this, but then work through the problem. Only code their mathematical cognitive processes.
- If they explicitly guess, code as such, but ensure that they did not make an informed guess based on earlier steps.
- Listen for key terms that imply testwise strategies may be in use (e.g., “plug in,” “eliminate,” or “ignore”). Code the processes appropriately to convey this implication.

Process for Calibration

1. The researcher and a second rater will independently listen to the audio files and view the transcripts and student work.
2. Both raters will code two students’ processes for each item (2 X 9 items = 18 codings).
3. The raters will compare their coding.
4. If discrepancies in coding are found, the raters will engage in a discussion about the process of and the reason for each of the codes.
○ The raters will agree on acceptable codes and terminology.
○ If necessary (i.e., if processes vary across raters), the raters will code student processes for five additional items.
5. The researcher will code the rest of the student processes, periodically rechecking calibration in order to preserve coding fidelity.