MASTERS, JAMES S., Ph.D. A Comparison of Traditional Test Blueprinting and Item Development to Assessment Engineering in a Licensure Context. (2010) Directed by Dr. Richard M. Luecht and Dr. Terry A. Ackerman. 117 pp.

With the need for larger and larger banks of items to support adaptive testing and to meet security concerns, large-scale item generation is a requirement for many certification and licensure programs. As part of the mass production of items, it is critical that the difficulty and the discrimination of the items be known without the need for pretesting. One approach to solving this need is item templating, an assessment engineering (AE) approach that is intended to control item difficulty and other psychometric operating characteristics for a class of items developed from each template. There are important advantages that can accrue to having exchangeable items that operate in a psychometrically similar manner in terms of item bank development (reduced time and lower cost to develop), pretesting efficiency, test security, and so forth.

This study describes one method to use AE and item templates in a licensure context to yield sets of items with statistical characteristics that match the needs of the program with reduced need for pilot testing. It is shown that item variants developed in this method fit the Rasch calibration/scoring model as well, if not better than items developed in traditional ways and that the item variants from the same template yield similar classical and IRT statistics. One key result of the study is a method to use AE to evaluate the performance of item writers over time.

A COMPARISON OF TRADITIONAL TEST BLUEPRINTING AND ITEM DEVELOPMENT TO ASSESSMENT ENGINEERING IN A LICENSURE CONTEXT

by

James S. Masters

A Dissertation Submitted to
the Faculty of the Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2010

Approved by

Richard M. Luecht
Committee Co-Chair

Terry A. Ackerman
Committee Co-Chair

APPROVAL PAGE

This dissertation has been approved by the following committee of the Faculty of the Graduate School at the University of North Carolina at Greensboro.

Committee Co-Chair _____
Richard M. Luecht

Committee Co-Chair _____
Terry A. Ackerman

Committee Members _____
Rick Morgan

_____
Scott Richter

_____
Betty Bergstrom

_____
Date of Acceptance by Committee

_____
Date of Final Oral Examination

ACKNOWLEDGEMENTS

I'd like to thank my advisor and committee co-chair Ric Luecht for assisting me with the basic premise and methodology behind this thesis, as well as insightful suggestions and recommendations along with encouragement in finishing the work. I'd like to thank Terry Ackerman, my first advisor and co-chair of my committee for his sage counsel throughout my years at UNCG. I would also like to thank Betty Bergstrom of Pearson VUE for participating on my committee, for providing me with the opportunity to use Pearson VUE resources and data, and for all of her encouragement over the last eight years. I thank the other members of my committee, Rick Morgan and Scott Richter for their time and advice without which this dissertation would not be what it is. I am also indebted to Nancy Thomas Ahluwalia and Bob Smith of Educational Testing Services who were my original mentors and who introduced me to test development and psychometrics. I'd like to thank Kirk Becker, Kathi Gialluca, Sue Steinkamp, and Pearson VUE for financial, moral, and intellectual support in graduate school and on this dissertation. I also thank Everett Kenyatta, Chris Beer, and Robin Ingalls of Pearson VUE for their assistance on the development and review of the items and templates used in the study. Thank you also to Kelly Kruger and Cindy Colonius of the Illinois Division of Insurance along with the other members of the Life and Health Committee (Dan Morgan, Ron Kotowski, Thomas Lubben, Len Karpowich, Chuck Budinger, Mike Teer, Bill McAndrew,

and Yvonne Clearwater) for assistance with reviewing the items and the content blueprint. Finally I thank my wife Leslie, who is my inspiration and has been my greatest support and encouragement since before I began working on this thesis.

.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

The primary purpose of this dissertation is to determine whether Assessment Engineering (Luecht, 2006b, 2007a, 2008b) can be used to develop simple task models and associated templates for extended-matching, true-false, multiple-choice test items (i.e., binary selected response items) in a licensure testing context. With the need for larger and larger banks of items to support adaptive testing and to meet security concerns, item generation is a very important topic (Bejar 2002; Downing 2006; Drasgow, Luecht, & Bennett, 2006; Embretson 2002; Hambleton & Jirka 2006; Luecht, Burke, & Devore, 2009; Raymond & Neustel 2006; Wainer 2002). Item templating is an assessment engineering (AE) approach that is intended to control item difficulty and other psychometric operating characteristics for a class of items developed from each template. There are important advantages that can accrue to having exchangeable items that operate in a psychometrically similar manner in terms of item bank development (reduced time and lower cost to develop), pretesting efficiency, test security, and so forth.

AE represents a relatively new area of research in psychometrics and educational measurement.

> AE represents a compendium of computer-science software and data systems design frameworks, engineering design principles, and psychometric technologies that focus on principled assessment task and instrument design in the service of providing useful and consistent measurement information for specific assessment purposes. (Luecht et al., 2009)

Figure 1 depicts the different components of an AE system. The left side of the figure shows a sample construct map, which is an ordered list of performance claims. In this example, the performance claims (Obscure Rules & Regulations, Advanced, Core Knowledge, and PreCore Knowledge) are different levels of knowledge of the rules and regulations associated with selling insurance. In the middle are a series of task models (each X represents a distinct task model) developed to measure knowledge at the specific location on the construct map. For example, the task models associated with Medical Supplements would measure knowledge somewhere between Core and Advanced. From each task model, multiple templates can be designed to generate multiple items.

In licensure testing, the test items are used to determine whether the candidate understands the rules and regulations associated with the profession, and whether he/she can distinguish true statements about the rules and regulations from false statements. Licensing exams “Typically deal with an applicant’s knowledge and skill at applying relevant principles, laws, rules, and regulations” (Shimberg, 1981). For this study, the extended-matching, true-false

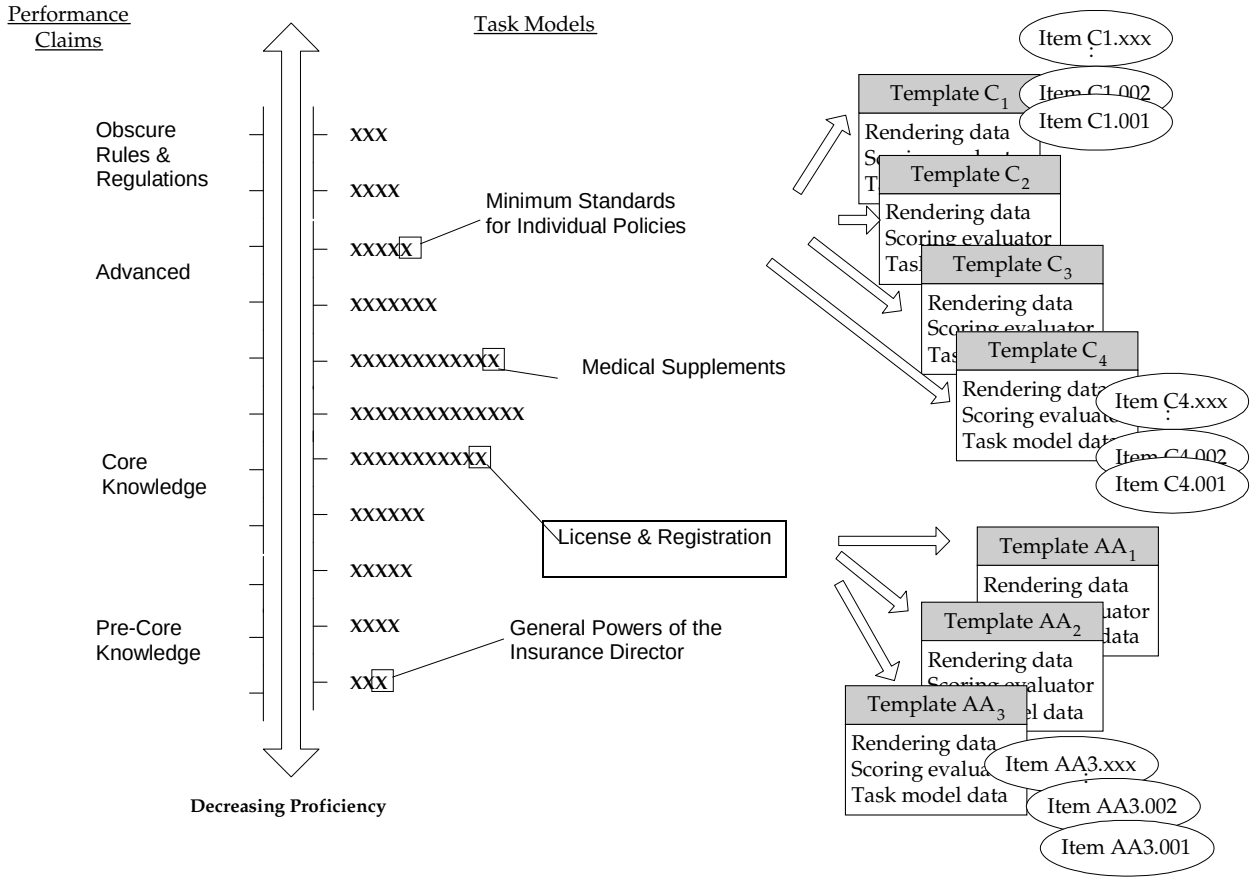


Figure 1. Assessment Engineering Components

(EMTF) multiple-choice (MC) item type is used to determine whether the applicant has the requisite knowledge. The items present a stem followed by a list of statements or propositions that are judged as "true" or "false". This is a simple declarative knowledge task in which the examinees must identify a positive or negative relationship between the primary object in the stem and the properties of each proposition. The following is an example of an EMTF item:

Limited lines producers are authorized to sell which of the following products?

(A) Medicare Supplement Policies
(B) Term Life Insurance
(C) Hospital Indemnity Insurance
(D) Industrial Life Insurance

The stem is “Limited lines producers are authorized to sell which of the following products?” Each of the four options (A, B, C, and D) is either true or false, and the test taker is expected to select the one option that is true. The key is C, which is the one true statement. The wrong options, A, B, and D, are the distracters. With three true options and one false option, it is also possible to ask for the one false option with a slight modification to the stem (e.g., “Limited lines producers are NOT authorized to sell which of the following products?”).

The process to develop the EMTF items began with Subject Matter Experts (SMEs) decomposing existing operational EMTF items into simple task models. The task models were intended to reflect varied levels of the complexity for declarative knowledge tasks on a licensure test. A relatively straightforward rating system was developed to allow the SMEs to rate the individual options; this rating system is explained further in chapter III. The task models and ratings were then used to develop true-false item templates with five to eight item variants generated per template. The template consists of:

- the topic and the topic difficulty rating,
- the question (stem),

- a list of possible true and false options (as many as possible), and
- a set of rules that can be used to develop multiple test items.

A sample template can be found in chapter III.

The construct map was developed using the exam blueprint which already existed for the given program. Each topic in the blueprint represents a topic that the entry-level practitioners (candidates who pass the exam) are expected to know and understand. The SMEs rated each topic based on the frequency that the entry-level practitioners would be expected to use the knowledge, and the importance that knowledge of the given topic would have to the performance of the job. The expectation is that the more frequently the knowledge is used and/or the more important the knowledge is to the performance of the job, the easier the questions about the topic would be.

Candidate response data was collected by assigning the items to pretest slots during a recent operational testing window. The data was subsequently used to evaluate the effectiveness of the templates to control difficulty and produce essentially exchangeable item variants.

Research Questions

Three research questions are being addressed in this dissertation.

1. Do the item variants developed from the templates fit a Rasch calibration/scaling model?

2. Will the item variants developed from each item template (and each associated task model) yield similar classical and IRT statistics?
3. Can logically determined item difficulty (item complexity scores) be used to replace or at least supplement item difficulty estimates computed using empirical data, and what is the effect on examinee scores?

Computer-based testing has driven the need for more and more test items.

> Continuous testing provides great challenges to testing programs. Perhaps foremost among these challenges is test security: Use of a single test form for an extended time period invites test compromise. Consequently, multiple forms or some type of adaptive testing must be used. This, in turn, means that many more items must be generated for the testing program. Technology—and an engineering approach—is critical for mass production of items. (Drasgow et al., 2006)

As part of the mass production of items, it is critical that the difficulty and the discrimination of the items be known without the need for pretesting. "If item difficulty and discrimination must be determined through pretesting, much of the potential flexibility and cost savings of automated item generation is lost" (Clauser & Margolis, 2006).

Definition of Terms

The following list defines terms relevant to this study and which will be used throughout the dissertation:

1. *Auxiliary Information* – Any additional material, in either the stem or option, required to generate an item, including texts, images, tables, and/or diagrams (Gierl, Zhou, & Alves, 2008).
2. *Cluster Scores* – Scores based on the sum of all options in an MTF item rather than the individual option scores when calculating test reliability.
3. *Enemy Items* – Two items that either are both testing the same concept or one gives a clue to the correct response of the other.
4. *Extended-Matching Item Type* – A multiple-choice variation that uses a long list of options linked to a long list of item stems (Haladyna, 2004).
5. *Incidental* – Features of an item that may be varied without an influence on the item performance (Clauser, Margolis, & Case, 2006). See also Radical definition in this list.
6. *Isomorphs* – Items generated with the constraint that they all be of the same psychometric attributes (Bejar, 2002).
7. *Item or Test Item* – The basic unit of observation in any test. A test item usually contains a statement that elicits a test-taker response (Haladyna 2004).
8. *Item Generation* – Any procedure that speeds up the item development process (Haladyna, 2004).

9. *Item Model* - An explicit representation of the variables in an assessment task. An item model includes the *stem, options,* and *auxiliary information* (Gierl et al., 2008).
10. *Item Options* – The alternative answers with one correct option and one or more incorrect options or distracters (Gierl et al., 2008).
11. *Item Shell* – The syntactic structure of a multiple-choice item. The item writer has to supply his or her content, but the stem or partial stem is supplied to give the item writer a start in the right direction (Haladyna, 2004).
12. *Item Templates* – A set of rules that can be used to develop multiple test items.
13. *Licensure* – "A process by which an agency of government grants permission to an individual to engage in a given occupation upon finding that the applicant has attained the minimal degree of competency required to ensure that the public health, safety, and welfare will be reasonably well protected" (U.S. Department of Health, Education, & Welfare, 1977, p. 4, as cited in Shimberg & Roederer, 1994).
14. *Multiple-Choice (MC) Item* – A multiple-choice item consists of two parts: A *stem* and a set of responses, one of which is correct (*key*) and the other(s), the *distracters*), wrong (Cantor, 1987). The

multiple-choice item can assume a variety of types, including absolutely correct, best-answer, and those with complex alternatives (Osterlind, 1998).

15. *Multiple True-False (MTF) Item Type* – Items that consist of a stem followed by several true-false options, each one of which must be responded to as true or false.
16. *Operational Items* – Operational items count toward whether a candidate passes or fails an exam.
17. *Pooled Pretesting* – The process where more pretest items are included with an exam than will actually be delivered to each candidate.
18. *Pretest Items* – Items that are administered to a candidate, but that do not count toward his or her score.
19. *Radical* – A term used to describe features of items that influence performance (Clauser et al., 2006). See also Incidental definition in this list.
20. *Rasch Model* – The model relating the ability of person *n* and the difficulty of item *i* to the performance of person *n* on item *i* is the objective model of measurement known as the Rasch model (Wright & Stone, 1999).

21. *Stem* – The part of an item which formulates context, content, and/or the question the examinee is required to answer (Gierl et al., 2008).
22. *Strong Theory (Item Generation)* – Calibrated items which are generated automatically using the design principles articulated in a cognitive model (Gierl et al., 2008).
23. *Subject Matter Expert* – A person who has substantial knowledge and experience in a particular field. In licensure testing, subject matter experts are used to validate the test items and the exams.
24. *Task Models* – Task models are used to define a unique combination of skills and knowledge objects required to support proficiency claims within a specific region of the construct-based measurement scale.
25. *Test* – A measuring device intended to describe numerically the degree or amount of learning under uniform, standardized conditions (Haladyna 2004).
26. *Variants* – In the terminology of automatic item generation, the abstract description the computer uses to generate instances of a class is called an item "model" and the instances are called "variants." (Drasgow et al., 2006). The term item variant is also used to refer to instances of an item model that range in difficulty or some other psychometric characterization of the items (Bejar 2002).

27. *Weak Theory (Item Generation)* - Calibrated items generated automatically using design guidelines (rather than design principles) discerned from a combination of experience, theory, and research (rather than cognitive models) (Drasgow et al., 2006).

CHAPTER II

LITERATURE REVIEW

The review of literature is presented in four parts. The first section presents an overview of traditional test and item development. Specifically, general test development principles starting with blueprint development and continuing with item development will be reviewed. In the item development section, different types of multiple-choice items will be described along with some basic item writing rules.

The second part of the literature review focuses on assessment engineering (AE) and automatic item generation (AIG). An overview of AE and the need for AIG is followed by a history of recent work in this area. AIG from strong theory is contrasted with AIG from weak theory, highlighting the advantages and disadvantages of each. Discussion about AIG includes the different philosophies that have been used to automatically generate test items.

The third part of the literature review focuses on methods to predict and model item difficulty. This section is divided into the different methods that have been used to predict item difficulty, an overview of some of the studies that have been performed and their results, and a section on calibrating the model rather than calibrating the item.

The last part of this chapter summarizes the first three parts in the context of licensure testing and the research questions addressed by the study.

Traditional Test and Item Development

Schmeiser and Welch (2006) stated, “The *Standards* (AERA, APA, & NCME, 1999) constitutes a seminal guide for proper test design and development” (p. 307). Chapter 3 of the *Standards* covers test development, which is described as “The process of producing a measure of some aspect of an individual’s knowledge, skill, ability, interests, attitudes, or other characteristics by developing items and combining them to form a test, according to a specified plan” (p. 37). Standard 3.6 explains that “The type of items, the response formats, scoring procedures, and test administration procedures should be selected based on the purposes of the test, the domain to be measured, and the intended test takers” (p. 44).

Downing (2006) described 12 steps for effective test development and provided a brief summary of tasks, activities, and issues; selected relevant *Standards* are noted (Table 1.1, p. 5). The steps in the test development process that are relevant to this study are (4) item development and (8) scoring test responses. Downing explained “The creation and production of effective test questions, designed to measure important content at an appropriate cognitive level, is one of the greater challenges for test developers” (p. 10). Included in the process are determining the item format (selected response or multiple choice

vs. constructed response) and training of the item writers or subject-matter experts.

The item format that is being examined in this study is the EMTF item format, which is a combination of the extended matching (EM) item format and the multiple true-false (MTF) item format.

The EM item format includes a list of stems and a list of options, all associated with a theme. The test taker is expected to select from the list of options the one that matches each of the stems, given the theme of the items. Normally the stems would be on the left side of the page and the options would be on the right side of the page. Here is a simple example showing the EM item format:

> For each of the calculations on the left side of the page, select the correct answer from the right side. Each answer may be used once, more than once, or not at all.
>
> | | | | |
> |---|---|---|---|
> | 1. | 9 + 7 = | a. | 13 |
> | 2. | 32 – 15 = | b. | 14 |
> | 3. | 4 x 5 – 6 = | c. | 15 |
> | 4. | (3 + 3) x 3 = | d. | 16 |
> | 5. | 4^2 = | e. | 17 |
> | | | f. | 18 |

Frisbie and Sweeney (1982) explained that:

> MTF items resemble MC items in their appearance. However, rather than selecting one best answer from several alternatives, examinees respond to each of several alternatives as separate true-false statements. These separate statements have a common lead or stem just like an MC item. (p. 29)

The MTF can be compared to the complex multiple choice item-type where candidates are expected to choose among options that consist of combinations of the TF statements. Here is a simple example of the complex multiple choice item format:

Which of the following statements about the topic are correct?

I. Statement A

II. Statement B

III. Statement C

A) I and II only

B) I and III only

C) II and III only

D) I, II, and III

Haladyna (2004) described the MTF format as “A viable alternative to the complex MC format” (p. 81). Frisbie (1992) stated “MTF items are consistently

more reliable than single-best response MC questions, when reliabilities are adjusted for equal amounts of testing time" (p. 22).

The EMTF item format has a stem, which asks which (one) of the options about a given topic is true (or FALSE), followed by a list of options, only one of which is true (FALSE). [Words in the stem in all caps are for emphasis.]

Case, Swanson, and Ripkey (1994) performed a study comparing the standard MC format with the EM format. EM sets with many (>5) options were used to create standard 5-choice MC items in two ways. In one situation, SMEs selected the four wrong options they believed to be the best distractors. In the second situation, statistics were used to select the four distractors that most candidates had selected. The different item types were administered to different test takers, where all test takers got some of each type. The results showed that the EM items were more difficult and more discriminating than the MC items with options chosen based on difficulty, and those items were more difficult and more discriminating than the MC items with options selected by the SMEs. The time required for a candidate to complete one of the EM sets is slightly longer than the time required for a 5-choice MC item, but when you group two items to the same set of options for the EM set, the time is comparable. The authors stated:

> This study demonstrated a strong relationship between the number of options used and the psychometric quality of the test material. Increasing the number of options had a marked influence on mean item difficulty, both because [of] a lower probability of guessing the correct answer and

> because the committee members were not always able to select the most functional distractors. (Case, Swanson, & Ripkey (1994, p. S2)

Frisbie and Sweeney (1982) stated:

> The MTF format would [also] appear to have several advantages over the MC format: a greater number of responses can be obtained in a given time period, the longer test is likely to be more reliable, a greater range of content can be examined because of the length, and more valid measure should be obtained because of the increased reliability. (p. 29)

The study by Frisbie and Sweeney showed that test takers can respond to 3.5 times more MTF items than MC items and MTF items show greater reliability, even when cluster scores are used[1].

Assessment Engineering and Automatic Item Generation

Bejar (2008) stated:

> In the last century test development was easy: items were cheap and did not need to be disclosed, specifications were loose, equating could be counted on to save the day, and validation was ad hoc. Today items are increasingly expensive, delivery systems are complex, tests have shorter life spans, and equating may not always be feasible. (pp. 2 & 3)

Ebel (1962) commented, "The process of test construction often appears to have more in common with artistic creation than with scientific measurement!" (p. 22). Cronbach (1970), in his review of *On the Theory of Achievement Test Items* (Bormuth, 1970), started by stating:

[1] For a test with MTF items, the reliability of the test can be calculated based either on the score of each option or based on the clusters of scores for all of the options for each MTF item.

> The design and construction of achievement test items has been given almost no scholarly attention. The leading works of the generation – even the Lindquist *Educational Measurement* and the Bloom *Taxonomy* – are distillations of experience more than scholarly analyses. (p. 509)

That being said, item development and item generation has come a long way in the last 50 years and it will progress much farther in the next 50 years when the promise of AE is realized.

AE is a new approach to test design, development, and delivery. Luecht (2007c) explained that AE "Provides an integrated framework with replicable, scalable solutions for assessment design, item writing, test assembly, and psychometrics" (p. 2). The primary steps that are included in AE are: (a) construct mapping, (b) evidence modeling, (c) task modeling and construct blueprinting, (d) template design and item writing, and (e) psychometric calibration and scoring. Each of these steps will be described.

Huff (2008) stated, "All AE approaches require articulation of student (candidate) expectations. The expectations must be characterized as observable evidence in order to be optimally leveraged for assessment design" (p. 2). The development of a construct map provides a description of what is known at different levels and what is used to articulate the expectations. The construct map should describe specific expectations at all levels on the scale. Figure 2 shows a simple example from 3rd grade science (Kennedy, 2008).

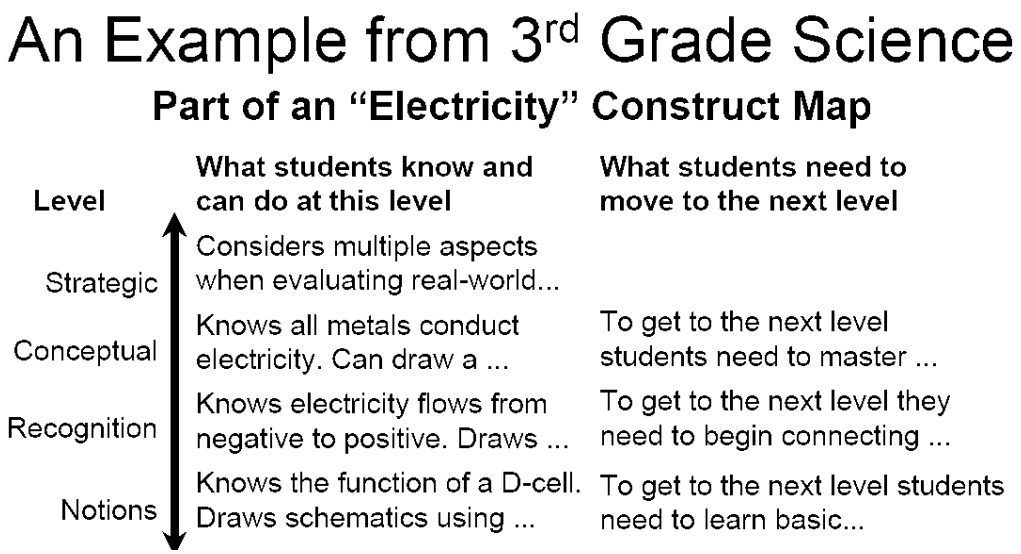


Figure 2. Simple Construct Map

Luecht (2007c) stated "Construct mapping amounts to clearly documenting a progression of ordered claims about proficiencies and skills and the required observable evidence needed to make those claims" (p. 13).

Luecht (2007c) went on to describe an evidence model as:

> A documented specification of the universe of tangible actions, responses, and/or products that would qualify as evidence for a particular proficiency claim. Each claim that is made about a test or assessment should have one or more evidence models to confirm the claim. The components of an evidence model include: valid settings or contexts, the plausible range of challenges for the target population, relevant actions that could lead to a solution, dangerous or inappropriate actions, legitimate auxiliary resources, aids, tools, etc. that can be used to solve the problem, and concrete exemplar products of successful performance. (p. 17)

For licensure tests, components would include knowledge of the rules and regulations associated with the license.

Task models are composed directly from the evidence models. Task models are used to define a unique combination of skills and knowledge objects required to support proficiency claims within a specific region of the construct-based measurement scale. Luecht (2007c) described the rules for building task models as:

1. Task models should be incremental–that is, ordered by complexity.
2. Task models at the same level must reflect conjunctive performance.
3. Higher performance assumes that lower level knowledge and skills have been successfully mastered. (p. 32)

Luecht also stated:

> Task models differ in location (difficulty) along the construct map. Each model provides measurement information in a particular region of the construct map. Deficits or gaps are filled by adding more task models. Ordering of task models can be empirically confirmed. (p. 27)

For each task model, multiple task templates can be constructed. Task templates are used by item writers to generate multiple test items with similar (exchangeable) information. The number of task models, task templates, and test items developed at different points along the scale is proportional to the needed measurement precision.

When describing AE, Luecht (2008a) explained that “Psychometric procedures are used as statistical quality assurance mechanisms that can

directly and tangibly hold item writers and test developers accountable for adhering to the intended test design" (p. 2). As evidence, Glas and van der Linden (2003) demonstrated that task models and/or templates can be calibrated instead of individual items, using a hierarchical Bayes framework. That means that one set of parameters is estimated for an entire family of items from a task model or template. The obvious advantages are (a) less pretesting, (b) robust parameter estimation, and (c) misfit is minimized if the families are well formed (Luecht, 2007c). Luecht (2009) explained that "Individual items *inherit* the estimated psychometric characteristics of the class via the task models and/or the templates" (p. 3).

History of Automatic Item Generation

There are different ways to classify the generative modeling systems that have been developed over the last 40 years. Bejar (2002) has developed a very useful, 3-level system. The lowest level in his hierarchy is the *functional* level. In the functional level, the emphasis is on generating items, but "Without explicit consideration of the constructs under measurement or a detailed modeling of responses" (p. 199). Examples include the work of the late 60s and early 70s, attributed to Hively, Patterson, and Page (1968) and Bormuth (1970).

Bejar's (2002) second level of generativity is called *model-based*. "At this level the generation of items is guided by models of performance, for example a cognitive analysis relevant to the domain under consideration" (p. 199). Three

examples of model-based generativity are the work of Enright and Sheehan (2002), Mislevy, Steinberg, and Almond (1999), and Hornke (2002), all of which are described in different chapters of *Item Generation for Test Development*, edited by Irvine and Kyllonen (2002). The work of Enright and Sheehan (chap. 5) uses cognitive theory to account for differences in item difficulty. Mislevy, Steinberg, and Almond (chap. 4) examine how the way that tasks and items are constructed affect the evidence that we can gather and the inferences we can make about scores. Hornke (chap. 6) uses literature or practical considerations to formulate Item Design Rules that help human item writers create a large set of items for each item type.

Models with the highest level of generativity are classified *grammatical* approaches by Bejar (2002). At this level "The item-generation and psychometric modeling are completely intertwined in such a way that it becomes possible to not only generate items, but also 'parse' any item to characterize its psychometric properties" (p. 200). Examples of this level of modeling can be found in the work of Bejar and Yocum (1991) and Embretson (2002). AE (Gierl, et al., 2008; Luecht et al., 2009) would also be classified as grammatical modeling.

As noted earlier, computer-based testing is driving the need for larger and larger banks of calibrated items. However, item generation predates computer-based testing. Hively et al. (1968) used item forms to develop items for a battery

of arithmetic tests. Their systems allowed for the development of an infinite number of mathematics problems. Each item form included fixed text for a mathematics word problem along with variable elements and rules used to take the place of these elements.

During the 1960s, the traditional methods of constructing achievement tests was for a test specialist to outline the content of classroom instruction, the cognitive behaviors the students should use to exhibit their mastery of the content, then write test items that hopefully tested the content at the appropriate cognitive level. There was no consistency in the items that were developed. Bormuth (1970) proposed a new methodology for developing achievement tests that operationalized the development of the test items. In Bormuth's method, everything was operationalized; the test writer had no options. Bormuth's research sought to develop general rules that would fully describe the relationships among the items, their responses, and the classroom instruction.

LaDuca, Staples, Templeton, and Holzman (1986) described a fairly complex method to generate test items to be used on medical tests. Their method involved decomposing existing items with good psychometric properties and identifying changes to the stem and resulting changes to the options. They explained:

> Using the source item as a model, doctor experts identify significant alternatives under each category of stem content, and stipulate the nature of the incorrect options (distractors). Completing the process leads to

> formulation of a set of item specifications for a family of test items addressing the same or closely related evaluative objective, and, in the extreme, exhaustive of the source item's major variants. (p. 53)

A taskforce that had been trained in the item modeling method was able to develop 10 item models, and then a knowledgeable nondoctor was able to use the item models to develop 100 acceptable items. LaDuca et al. went on to state that "The task force agreed that a review of specifications might serve as a satisfactory substitute for review of the items themselves" (p. 56). While the task force estimated that the items may be equivalent, no empirical evidence was available as of the publication of the article.

The work of Bejar (1990, 1993, 1996, 2002) and Bejar and Yocum (1991) is characterized as generative modeling. Bejar and Yocum described generative modeling as consisting "Of encoding information about the cognitive processes and structures that underlie test performance into an item-generation algorithm in such a way that the generated items have known psychometric parameters" (p. 129). They explained that "Generative psychometrics, involves a 'grammar' capable of (a) assigning a psychometric description to every item in the universe of items, and (b) generating all the items in the universe of items" (p. 130). Bejar (1993) gave examples of generating items for spatial ability, reasoning tests, verbal ability, and complex skills.

Meisner, Luecht, and Reckase (1993) generated mathematics test items using algorithmic methods. They hypothesized that items "Requiring the same

knowledge and skills for their solution would be likely to exhibit parallel statistical characteristics for a given population of examinees" (p. 6). For this study, they used item forms (Hively et al., 1968) to create algorithms to produce pretest items for eight forms of the ACT Assessment Program Mathematics Test. The results were very promising for some of the sets of generated items, and explanations could be hypothesized for the differences in the other sets.

Robert Mislevy is credited with the development of Evidenced Centered Design (ECD). Evidence-centered assessment design is an approach to constructing educational assessments in terms of evidentiary arguments (Mislevy, Almond, & Lukas, 2003). The assessment process with ECD begins with domain analysis where substantive information is gathered that has implications for the assessment. The next phase is the modeling of the domain by determining the knowledge, skills, and abilities associated with the domain, and identifying the elements that need to be included in the assessment. The assessment arguments are turned into items and tasks through a *conceptual assessment format*, which includes three related models: (a) the student model, (b) the evidence model, and (c) the task model. The student model defines *what* will be measured, the evidence model defines *how* it will be measured, and the task model defines *where* the elements will be measured. ECD also includes a *4-process architecture for delivery*: (a) tasks are selected, (b) tasks are

presented, (c) responses are collected, and (d) responses are scored. The architecture for delivery allows for continuous updating of items and parameters.

Susan Embretson developed a process she calls the cognitive design system to generate items. Implementing the cognitive design system approach involves studying the cognitive components of item solving prior to test development (Embretson, 1999). Through 1999 Embretson had examined verbal analogies, verbal classifications, geometric analogies, geometric classifications, series completion, paragraph comprehension, spatial folding, mathematics word problems, matrix completion problems, and spatial construction items and found that she was able to predict item psychometric properties at least moderately well. Gorin and Embretson (2006) reviewed reading comprehension items and were able to show some relationship between the cognitive features of the items and item difficulty, but work still remains. Two issues that could affect the psychometric properties are the relative proportion of variable elements in the item structures and the differences between the substituted elements (Embretson, 2002). One of the major advantages of the cognitive design system approach is that construct validity is assessed at the item level.

Enright and Sheehan (2002) used cognitive theory to both analyze the difficulty of word problems from the quantitative sections of the Graduate Record Exam (GRE) and to examine an augmented item classification system for the

quantitative sections of the GRE. A series of studies is described in which difficulty modeling is used to clarify the constructs assessed by quantitative items on graduate admissions tests in order to develop a more principled basis for item generation. They had good success with rate problems (manipulated features accounted for 90% of the variance in difficulty), but less success with probability problems.

Newstead, Bradon, Handley, Evans, and Dennis (2002) used the psychology of reasoning to develop two methods to generate items for specific tasks. The first method involved developing items for a spatial test (directions and distances) used in officer selection for the British Royal Navy. The second method involved developing analytical reasoning items (literacy and numeracy) for Educational Testing Service. For the first approach, "The item universe is defined as the total set of items that can be produced by the factorial combination of a specified set of features and their levels" (p. 57). The group was relatively successful at generating items and predicting the difficulty with some accuracy for this item type. The primary limitation to the first approach identified by the group is that the "generative framework may be decoded by somebody with access to a moderately sized sample of items that it has produced" (p. 63). The second approach deals with logical reasoning items. Given a scenario, candidates can be asked what must be true (*necessity item*), what is possible to be true (*possible item*), which option could possibly be true (*possibility item*) or

which option is not possible (*impossibility item*). Dennis et al (2002) explained that with this approach the group did not have as much success predicting item difficulty due to the "combined complexity of the rules making up the initial rule set and of the stem rule" (p. 65).

Singley and Bennett (2002) applied schema theory to the generation of items for a mathematics assessment. They explained:

> The key to this theory is the basic assertion that math problems can be characterized (and categorized) in terms of the underlying set of equations that relate the entities of the problem to one another. According to this analysis, problems that superficially appear quite distinct may in fact be instances of the same underlying problem structure, or *schema*." (p. 361)

Taxonomy of Item Modeling

Gierl et al. (2008) provided a taxonomy of item modeling. They suggested that item modeling techniques vary based on the logic included in the stem and the logic included in the options. There are four possibilities for the stem: (a) independent, (b) dependent, (c) mixed, or (d) fixed. Independent implies that a change in one element of the stem has no effect on other elements of the stem. Dependent means that a change in an element of the stem directly affects other elements. Mixed means that the stem includes some elements that are independent and other elements that are dependent. A fixed stem is the same for all variants developed from the model. The options can be either randomly selected from a pool, constrained according to formula, calculation, or context

based on the elements in the stem, or invariant. This information is summarized in Table 1.

Table 1

Taxonomy for Item Modeling

	Options		
Stem	Randomly Selected from a Pool	Constrained (generated according to formula, calculation, or context)	Invariant
Independent	√	√	√
Dependent	√	√	√
Mixed Independent/Dependent	√	√	√
Fixed	Extended-Matching True-False	N/A	N/A

Note: From Developing a Taxonomy of Item Model Types to Promote Assessment Engineering by M. J. Gierl, J. Zhou, and C. Alves, 2008, *The Journal of Technology Learning and Assessment, Volume 7*(2). Copyright 2008 by M. J. Gierl, J. Zhou, and C. Alves. Adapted with permission.

The EMTF item template falls into the fixed stem/options selected from a pool category. That is, the stem is invariant with respect to the options—it always asks for the one true statement or the one false statement. The options are selected from a set of true and false statements associated with the stem.

It does not make sense to have a fixed stem and options that are either constrained by elements in the fixed stem or invariant. Gierl et al. (2008) gave

examples of each of the other cells in the taxonomy in their paper, but they will not be described here.

Another way to split up the different methods of item generation is by methods that are based on strong theory versus those based on weak theory. Drasgow et al. (2006) explained the distinction between item generation from strong versus weak theory. "The goal of automatic item generation from strong theory is to generate *calibrated* items automatically from *design principles* by using a *theory* of difficulty based on a cognitive model" (p. 474). With weak theory, the design principles are based on decomposing a "parent" item whose psychometric characteristics are known and using the "theory" of *invariance* (p. 474). The cognitive model in strong theory provides a detailed description of the variables that affect examinee performance which, in turn, can help pinpoint the item difficulty features (Gier et al., 2008). With weak theory, subject-matter experts decompose existing exemplar operational items and attempt to determine the aspects of the item that will and will not affect the difficulty of the items generated. Those aspects that will not affect item difficulty can be varied in a systematic way to create new items.

Examples of item generation from weak theory include the work of Hively et al. (1968), Bormuth (1970), LaDuca et al. (1986), and Meisner et al. (1993). Examples of item generation from strong theory include item generative modeling (Bejar, 1990, 1993, 1996, 2002; Bejar & Yocum, 1991), evidence

centered design (Mislevy, 2006; Mislevy et al., 2003; Mislevy & Haertel, 2007; Mislevy & Riconscente, 2005; Mislevy, Steinberg, & Almond, 1999), the work of Enright and Sheehan (2002), the cognitive design system approach (Embretson, 2002; Gorin & Embretson, 2006), the application of schema theory to mathematics assessment (Singley & Bennett, 2002), and assessment engineering (Gierl et al., 2008; Luecht, 2002, 2003, 2007a, 2007c; Luecht et al., 2009).

Predicting Item Difficulty

As Clauser and Margolis (2006) explained:

> If automated item generation is to reach its potential, it will be necessary to produce not only quality items but also items with known performance characteristics (if item difficulty and discrimination must be determined through pretesting, much of the potential flexibility and cost savings of automated item generation is lost). (p. 301)

There are many good reasons to attempt to determine or predict the statistical characteristics of items that are being developed through AIG. These include:

- Test developers are more able to construct statistically parallel forms (Meisner et al., 1993).
- It eliminates the need to pretest items or reduces the number of test takers required for efficient calibration of the items (Bejar, 1993; Chalifour & Powers, 1989; Frase et al., 2003; Gorin & Embretson,

2006; Hambleton & Jirka, 2006; Luecht, 2006a, 2006b; Meisner et al., 1993; Swygert, Scrams, Thompson, & Kerman, 2006; Wainer, 2002).

- Unacceptable items (too easy, too difficult, poor discrimination) are not developed (Luecht, 2006a, 2006b; Swygert et al., 2006).

Swygert et al. (2006) provided added insight into the value of AIG:

> One reason items are so expensive is that, once created, they must be pretested on a group of test takers similar to the test takers who will see the items operationally. This pretesting phase is necessary so that unacceptable items (e.g., items that are easily guessed, items that do not help discriminate among test takers, items that show differential functioning, etc.) may be identified and not used operationally. If a prediction model existed that showed how a certain type of item clone would behave, the characteristics of the item could be known with some degree of certainty ahead of time, and perhaps the number of test takers required to pretest the item could be reduced. (p. 1)

The greater the number of candidates who see items during field testing, the greater the risk of item compromise (Hambleton & Jirka, 2006). Wainer also provides insight:

> Obviously, if items were to be generated automatically and used within traditional printed test forms we could scrutinize the items in the traditional way. But subjecting them to content and sensitivity review as well as pretesting would mean that very little of the $1,000+ cost per item would be saved. The savings would accrue only if all of those steps could be omitted without worrying about the quality of the item. (p. 301)

Methods to Predict Item Statistics

Several of the methods that have been developed to predict a generated item's psychometric characteristics have already been discussed. Hambleton

and Jirka (2006) have developed a set of factors to consider when using judges or SMEs to estimate item statistics: “Effective training requires practice in rating items and receiving some feedback about how well they are doing” (p. 409). However, the methods that have shown the most promise are those that are based on accessing the cognitive processes and structures of the item models (Bejar, 1990, 1993, 1996, 2002; Bejar & Yocum, 1991; Chalifour & Powers, 1989; Embretson, 2002; Enright & Sheehan, 2002; Gierl et al., 2008; Gorin & Embretson, 2006; Luecht, 2002, 2003, 2007a; Luecht et al., 2009; Mislevy, 2006; Mislevy et al., 2003; Mislevy & Haertel, 2007; Mislevy & Riconscente, 2005; Mislevy et al., 1999; Singley & Bennett, 2002).

Studies

There have been many studies that have attempted to determine item statistics with little or no field testing. In addition to their own study, Hambleton and Jirka (2006) cited 18 other studies that used the judgment of SMEs to estimate item difficulty. Their analysis showed that the studies fell into five categories: (a) studies on judging item difficulty, (b) studies on other item characteristics, (c) studies on item-writing rules, (d) research on other attributes affecting item characteristics, and (e) a mixture of judgment and factor studies. In addition to those studies, there are also studies where the author(s) have

attempted to analyze the cognitive features of the items and determine which features impact the difficulty and which ones do not and how they interact.

Hambleton and Jirka (2006) listed 12 factors that should be considered when designing studies to estimate item statistics. Following is a summary of that list.

1. With training, judges do a more accurate job of estimating item statistics.
2. Component models of item difficulty (e.g., estimating item reliability, suitability of distractors, consistency of items with the item writing rules) are not typically as good predictors of item difficulty as more global ratings (i.e., estimate the level of difficulty of this item).
3. In many studies, raters are asked to use unfamiliar scales in judging item difficulty. These scales are often problematic for judges. Anchor-based rating scales appear to be more useful to them.
4. Predictors of item difficulty (factors that influence item difficulty) vary as a function of the item type. Effective training may need to consist of two aspects: Generic training (e.g., item readability is important, or the number of steps required to complete a problem is important), and specific training for particular item types.

5. Securing judge agreement about item difficulty provides more accurate estimates of item statistics than averaging totally independent judges' rating.
6. Item complexity impacts on the accuracy with which item statistics can be estimated.
7. Details associated with the test administration can be influential in estimating item difficulty.
8. Judges need to be trained to look at (a) structural characteristics, (b) surface features, and (c) the psychological component. The common shortcoming of judges is that they focus their judgments of item statistics on only one of these dimensions.
9. The candidate population is important. Judges either need to know the candidate population well, or time must be spent in training the judges to ensure that they have detailed information about the candidate population.
10. Item placement in a test is important.
11. There is considerable evidence to suggest that item difficulty levels can be predicted, but predicting item discrimination has been much more difficult to do with any accuracy.

12. Average ratings of item difficulty statistics across judges are much more highly correlated with actual item difficulty than individual judges' ratings.

Dudycha and Carpenter (1973) compared the difficulty of items with certain characteristics to items with other characteristics. The results showed that open-stem items (sentence completion) were more difficult than closed-stem items; negatively-worded items were more difficult than positively-worded items; and items with inclusive options (e.g., all of the above, or none of the above) were more difficult than items with all specific options. There were no interactions among the factors. Item discrimination was only affected by the inclusive versus specific option factor, with the items that had one or both of the inclusive alternatives having significantly lower discriminative ability.

Belov and Knezevich (2008) developed a process to predict item difficulty using semantic similarity measures based on a lexical database. Their method was able to improve the prediction of main-point reading comprehension items used in the Law School Admissions Test over other similar methods.

Rupp, Garcia, and Jamieson (2001) attempted to model item difficulty of reading and listening comprehension items as a function of 12 text and item/text interaction predictor variables. They found that 7 of the 12 variables accounted for 31% of the variance in item difficulty. They went on to use a nonparametric technique to uncover linear dependencies among the predictor variables, which

showed 7 of 12 variables (not the same 7) were relatively important. The two methods combined "provided a richer picture of the interrelations of variables that affect item difficulty" (p. 185).

Mislevy, Sheehan, and Wingersky (1993) attempted to predict item psychometric characteristics based on various sources of collateral information: Expert judgment, test specifications, and cognitive processing requirements. Thirty collateral variables were proposed by a team of test developers and two test developers coded all items on eight forms of the Pre-Professional Skills Test (PPST). The results of the study showed that:

> While there have been many advances in statistical methodologies, and studies have shown that collateral information can be predictive of item operating characteristics, at this point [1993] the information is not sufficiently rich to eliminate or substantially reduce pretesting and equating. (p. 76)

Calibrating the Model

One of the benefits of AIG is the ability to use smaller samples of pilot test takers to determine the calibrations of the items. As Bejar (1990) stated:

> It may be feasible to obtain valid estimates of difficulty by combining information about the psychological demands of items using a small sample of examinees, instead of administering the test to a large sample of potential examinees. The implementation of this approach would require procedures for estimating the parameters in a psychometric model that are capable of incorporating "prior" information into the estimation process. (p. 238)

Glas and van der Linden (2003) developed a multilevel item response model to deal with differences between the distributions of item parameters of families of item clones. Bejar et al., (2003) discussed calibrating the item model using an expected response function (ERF). They explained that "ERF is a procedure for attenuating parameter estimates as a function of the uncertainty in them" (p. 11). The ERF is used to account for uncertainty due to departure from isomorphicity among the variants that are pretested to calibrate the model (Drasgow et al., 2006). Glas (2006) described and compared two procedures that can be used to calibrate cloned items, a process the paper refers to as item clone modeling (ICM). Both methods use a Bayesian procedure for parameter estimation using a Markov chain Monte Carol (MCMC) method.

Bejar et al. (2003) designed a simulation study "to explore the effect an on-the-fly test design would have on score precision and bias as a function of the level of item model isomorphicity" (p. 2). That is, they wanted to determine how the items generated from an item model would affect test taker scores depending on how close the statistics of the generated items were to the predicted statistics. They found that under certain circumstances there was no bias, but measurement precision was reduced.

The Role of Item Generation in Licensure Testing

In the United States, licensure testing is used by states to determine whether a candidate is qualified to work in a given profession. Haladyna (2002)

stated "*Licensure* is a complex system of government regulation with the purpose of providing public protection" (p. 119). AERA, APA, & NCME (1999) explained that "Licensing requirements are imposed by state and local governments to ensure that those licensed possess knowledge and skills in sufficient degree to perform important occupational activities safely and effectively" (p. 156). There are many more definitions of licensure and licensure testing, but these two capture the essence – the purpose of licensure testing is to protect the public (Clauser et al., 2006; Downing, 2006; LaDuca, 1994; Schimberg & Roederer, 1994; Schmitt, 1995; Spray & Huang, 2000).

Wainer (2002) asked two questions: (a) "When do we need automatic item generation?" (p. 299), and (b) "When are computerized tests sensible?" (p. 300). While AIG can be useful in almost any situation, it is always necessary when we need computerized tests. One of the three answers Wainer gives to when computerized tests are sensible is "When the test results are needed year round, (e.g., licensing tests) and delays in testing yield concomitant delays in the examinee being able to earn a living" (p. 301). Therefore, for licensure testing, as with all high-stakes computer-administered tests, automatic item generation is needed.

To determine the item format that is most appropriate, it is important to determine whether the knowledge being tested is concrete or abstract (Haladyna, 2004). Shimberg (1981) stated that licensing exams "Typically deal with an

applicant's knowledge and skill at applying relevant principles, laws, rules, and regulations" (p. 1140). That is another way of saying that much of the knowledge being tested in a licensure exam is concrete, which requires low-inference item formats such as the EMTF item format.

Experience shows us that the pass rate on the Illinois Accident and Health Insurance State exam ranges between 75 and 85 percent of the examinee population. To maximize the reliability of the classification of a candidate as passing or failing, Shimberg (1981) stated "The test should include a large number of questions that should be answered correctly by *that percent* of the applicants" (p. 1142). Therefore, our goal should be to generate items in that difficulty range.

CHAPTER III

METHODS

The Methods chapter is divided into four parts: (a) Construct Map Development; (b) Template and Item Development, Item Pretesting, (c) Template and Item Calibration; and (d) Statistical Analyses.

Construct Map Development

As noted earlier, the primary purpose of a construct map is to anchor the evidence models, and position task models and item templates with respect to required measurement information. For the purposes of this study, the construct map was developed using an informal practice analysis. A similar process was used by Luecht (2008b) to build evidence models for an operational licensure program. A group of SMEs reviewed each of the topics in the existing test blueprint and rated them on a frequency scale and a criticality scale from the perspective of the entry-level practitioner. The expectation is that knowledge of topics used frequently will be more familiar to the entry-level practitioners, and therefore questions about those topics would be easier than topics where the knowledge is used less frequently. Similarly, knowledge that is more critical to acceptable performance of the job for the entry-level practitioner would also be more familiar, and therefore questions about those topics would be easier than

topics where the knowledge is less critical. Each topic was rated on a 4-point scale for both frequency and criticality. The SMEs were asked to use the following scales:

How frequently will the entry-level practitioner use the knowledge for the given topic on the job?

1. Daily
2. Weekly
3. Monthly
4. Less frequently

How critical is knowledge of the given topic to appropriate performance of the job?

1. Essential
2. Important
3. Useful
4. Less important

Template and Item Development

The templates and items were developed for the state portion of the Illinois Department of Insurance Accident and Health exam, which is developed and delivered by Pearson VUE. Illinois provides licenses to candidates who wish to sell insurance in Illinois. In licensure testing, many of the questions are used to determine whether the candidate understands the rules and regulations associated with the profession and whether he/she can distinguish true statements about the rules and regulations from false statements.

The Illinois Accident and Health exams were built to meet a content outline. Assisted by an insurance SME from Pearson VUE, the development of

the templates began with a gap analysis to determine which parts of the content outline were most in need of new item development. Fifteen different domains were identified as needing additional items. For each selected domain, a construct identifier was selected to narrow the focus of the template. Each template was referenced to the specific part of the Illinois State Insurance Code. The construct identifier was used to help develop the stem, which generally took the form of "Which of the following statements about *construct identifier* is **CORRECT**?" or "Which of the following statements about *construct identifier* is **FALSE**?"

The templates were developed in these areas by decomposing existing operational items that had good statistics and that had been recently administered to a representative sample of candidates. The task models were evaluated by the Pearson VUE SME to determine their complexity and the knowledge objects that were included. Next, the regulations associated with each template were reviewed in order to find true statements and false statements that could be used for the item options.

In addition to the construct map, which provides a way to evaluate the difficulty of the task models and item templates, the difficulty of the options was evaluated. For each option, true or false, the Pearson VUE SME rated the option on a 3-point scale as: simple (1), moderate (3), or complex (5), from the perspective of how likely a candidate would be to think the true statement was

false or the false statement was true. The less likely the candidate would be to assess whether the option was true or false, the higher the rating. An option is complex if both weak and average candidates think it is true (false) when it is actually false (true), while strong candidates correctly recognize that it is true or false. An option is simple if only very weak candidates will be unable to correctly determine whether it is true or false. An option is moderate if it is somewhere between complex and simple.

An insurance SME from the state of Illinois, who is part of the committee that works with Pearson VUE on the Illinois Insurance Exams, performed the final review of the templates. As part of the review, the difficulty of each option was assessed on the same 3-point scale that Pearson VUE SME had used. To determine the option difficulty, the two ratings were averaged.

The next step in the process was to determine which items in the existing bank of approved items were enemies of the different items from the given template. Two items are enemies if one gives away the answer to the other, or if they are both basically testing the same concept. Two items that are enemies of each other should not be administered to the same examinee during a single testing session. To determine which items were enemies, all items in the existing bank that used the same construct identifier were identified. Then for each approved item, the options that were the same (or close to being the same) as one of the options in the template were identified. Any items from the template

that had one or more options in common with the approved item were made enemies of that item.

After the items had been developed, the full committee reviewed the items.

Each template also includes a process that explains how to develop different items by analyzing the number of true and false options and their ratings, consisting of the position of the key and the total sum of the ratings of the options. In order to assure that the position of the key did not affect the item difficulty, the position was always the same for all items within a given template. As noted in chapter I, the difficulty of a given item is a function of the topic and the difficulty of the options. The actual item difficulty will be determined through pretesting. An exemplar template and several sample items developed from the template are shown in Appendix A.

Although 15 different templates were developed, only are 14 were included in the study. Table 2 summarizes information about the templates that were used. For five of the templates, the sum of the ratings was the same for all items developed from the template. For the others, there were either two different sums or the option sum varied across items. One of the item templates (9) did not fall into the EMTF item category, but follows similar item templating rules and has been included in the study. For this template, the options are fixed and the stem is Mixed (see Table 1).

Table 2

Summary of Template Information

Template	Number of Items	Item Type	Options Developed		Key	Sum of Option Ratings
			True	False		
01	5	Key is True	15	13	B	13
02	8	Key is True (5) Key is False (3)	8	15	D	10
03	8	Key is True (3) Key is False (5)	21	9	A	11.5 - 13.5
04	3	Key is True	4	11	D	15 or 16
05	4	Key is True	7	5	B	13 - 15
06	8	Key is True (5) Key is False (3)	5	9	C	13
07	5	Key is True	7	7	A	10 or 11
08	5	Key is False	16	13	B	9 - 13
09	5	Key is True				3 or 5
10	6	Key is True (4) Key is False (2)	6	18	A	13 or 15
11	5	Key is True	8	16	C	12 - 14
12	5	Key is True	6	8	D	10
13	6	Key is True (3) Key is False (3)	17	7	A	11 - 14
14	8	Key is True (5) Key is False (3)	12	13	B	11 or 12

A total of 89 pretest items were published. However, these items had to be published before the full committee had a chance to review them. While the

items had been reviewed by one Illinois SME and two different Pearson VUE SMEs, the full committee at their meeting (after the items had been published) identified 8 items that had significant issues (conflicting regulations, two correct responses, no correct response, or not appropriate for the candidate population). These items have been dropped from the analysis.

Item Pretesting

The Illinois Insurance examinations use pooled pretesting. For each operational form, there is a pool of pretest items (N>n where N is the number of pretest items in the pool and n is the number of pretest items that are delivered to each candidate). That means, in addition to the fixed set of operational items that count toward whether each candidate passes or fails, he or she is randomly delivered a subset of the pretest items available which do not count toward whether he or she passes. Three new operational forms of the Accident and Health Exam are placed in the field each year, and each has a pool of pretest items from which each candidate is randomly administered eight. The three pretest pools can overlap (have items in common). Appendix B shows the layout and the overlap of operational items within the three forms. Appendix C shows the same thing for the pretest items.

The Pearson VUE test driver for the Illinois Insurance exams is designed so that no two pretest items that are identified as enemies of each other will be administered to the same candidate. All pretest items developed from a single

template have been identified as enemies of each other. Therefore, each candidate will see, at most, one item from a given template.

Approximately 800 candidates take the Illinois Accident and Health exam each month. Each pretest pool is seen by approximately one third of the candidates and each pretest pool has 57 items. That means that each pretest item in a given pool will be seen, on average, by about 37 candidates per month. Most items are in two pools, but even items in only one pool should be seen by over 100 candidates in the five months that the exams are available. A sample size of 100 or more candidates for each pretest item is sufficient to give acceptable IRT (Rasch) and classical item parameters. (Depending on the difficulty of the item, the error associated with the calibration of the item based on 100 candidate responses is approximately 0.25 logits).

The pretest items could not be reviewed by the full Illinois Accident and Health Insurance Committee until after the items had been published. When the committee reviewed the items, problems were found with some of the options. These problems, for the most part, are minor, but they may require that the items be revised and repretested before the items are used operationally. However, the item statistics should still be relevant for these items. In the case of one template, there are conflicting regulations, so it is likely the statistics will not perform as expected, and it is likely that the items associated with this template will not be used. In an ideal situation, the templates would have been reviewed

by the full committee before the items were developed and published. The templates with problem items were not included in this study unless they still had at least five items without problems.

Template and Item Calibration

Data will be collected for five months. Each pretest item will be seen by between 100 and 300 test takers. The response data will be used to calibrate the item difficulties. In addition to calibrating the difficulty of each individual item, it will also be possible to calibrate the template difficulties using Winsteps[2] by combining the responses to all items for the given template. In order to make certain that all item calibrations are on the same scale as the operational items in the bank, the operational items will also be calibrated, but their difficulties will be held constant at the bank value. For each item and for each template, the following statistics will be calculated: difficulty (logits), error (logits), mean square infit, mean square outfit, point-measure correlation. Mean square infit and outfit are fit statistics calculated by Winsteps (Linacre, 2009). The point-measure correlation is the correlation of the candidates' performance on the item (or the template) with their ability.

Statistical Analysis

There are three research questions addressed by this dissertation.

[2] Linacre, J. M. (2009). Winsteps 3.69.0 [Computer Software]. Chicago, IL: Mesa Press.

1. Do the item variants developed from the templates fit a Rasch calibration/scaling model?

This question assumes that *current item-writing practices* form a baseline of sorts insofar as scaling the assessments. The use of templates and principled item development is intended to reduce (not increase) method variance or other subtle sources of scale contamination. Winsteps provides two different fit statistics based on the Chi Square distribution, Infit and Outfit. The Outfit statistic is the average of the squared standardized residuals times the information and Infit weights the observations by their statistical information. Since information or model variance is higher at the center and lower at the extremes, the Infit statistic is less influenced by outliers than the Outfit. Both statistics should have a mean of 1.0. I compared the average and standard deviation of Infit and Outfit for the items variants developed from templates to the operational items that were developed in traditional methods. I also provided plots that demonstrate the fit and variation among the item difficulty estimates.

2. Will the item variants developed from each item template (and each associated task model) yield similar classical and IRT statistics?

In the ideal situation, each template is expected to fully control item difficulty, making the items completely exchangeable. The practical utility and cost savings of having many exchangeable items that function in a psychometrically consistent manner is obvious. In that regard, this question

addresses the expectation that variation in item difficulty within each template will be less than difficulty variation across (between) templates/task models. For each template I compared the difficulties of the item variants using both IRT and classical analyses. The difficulty of each template was also determined by collapsing the data across the item variants from that template. I also compared the ratings as determined by the option ratings developed by the SMEs.

3. Can logically determined item difficulty (item complexity scores) be used to replace or at least supplement item difficulty estimates computed using empirical data and what is the effect on examinee scores?

Using item response theory (IRT), we *calibrated* items to estimate the difficulty of each item relative to an underlying proficiency scale. In the present context, five possible calibrated item difficulty statistics can be computed for each item: (a) item difficulty parameters directly estimated from empirical data using an IRT Rasch model, (b) item difficulty estimates computed from SME ratings of task complexity and option difficulty, (c) template difficulty ratings where the data are collapsed across all item variants, (d) template difficulty ratings where the data are collapsed across a subset of the item variants or (e) hybrid item difficulty estimates that augment the SME task complexity ratings with empirical data. When these five different item difficulty indicators are used in scoring the same examinees, how do their scores compare? If the logical ratings or hybrid

methods produce very similar scores to the empirically determined estimates, there is potential for greatly reducing the need for large examinee samples for item calibration purposes.

CHAPTER IV

RESULTS

A key purpose of this study involves the practical utility of devising templates to reduce the item exposure risks and costs in terms of examinee time and effort associated with pretesting every item. That is, if item templates help control the difficulty of items, and if the item difficulty estimates can be analytically computed (versus empirically estimated for every item), there are potentially enormous cost-reduction benefits that could be realized by testing organizations. The results are presented in this chapter in four sections: (a) an evaluation of model fit; (b) similarity of item statistics within template (class); (c) utility of logically determined difficulties based on content, and a consideration of item surface features; and (d) impact of calibration strategies on examinee proficiency scores.

Rasch Model Fit of the Item Variants

Model fit is an important aspect of devising a test scale and calibrating the item difficulties relative to that scale. Under the Rasch item response theory model (Rasch, 1960/1980; Wright & Stone, 1979), two parameters define the model for responding to dichotomously scored items: A person proficiency trait

parameter, θ, and an item difficulty, b_i, (i=1,…,n). The Rasch model can be expressed as a probability of obtaining a correct response:

$$\Pr\left(u_{i=1}\left|\theta_j, b_i\right.\right) \equiv P_{ij} = \frac{\exp\left(\theta_j - b_i\right)}{1+\exp\left(\theta_j - b_i\right)}. \qquad (4.1)$$

Model fit is essentially a function of the difference between the observed response data, u_{ij} (the dichotomous scores of zero or one for j=1,…,N persons and i=1,…n items) and the Rasch model probability, P_{ij}. One part of the residual $\varepsilon_{ij} = u_{ij} - P_{ij}$ is due to random error and is ignorable. Any remaining residual, especially that portion of which might covary with the residuals from other items, is termed misfit and may denote item scoring dependencies, multidimensionality, method variance, or other nonrandom sources of variance. The two commonly used fit statistics are termed Infit and Outfit (Wright & Stone, 1979). These two statistics differ primarily in the normalization method (denominator) used to conceptualize the [random] error variance. Infit uses an aggregate normalizing term and is considered to be most sensitive to misfit that occurs for items well-matched to each examinee's apparent proficiency (i.e., where $\theta_j - b_i$ approaches zero and P_{ij} approaches .5). Outfit uses a localized normalizing term and is typically sensitive to item–person misfit where items are located further away from the examinee's proficiency. Of the two statistics, Infit is generally considered to be more relevant since it signals potential misfit for those items

best suited for an examinee at a particular proficiency level. Both statistics are computed by Winsteps (Linacre, 2009) as part of the routine calibration outputs.

The Winsteps mean–square (MS) Infit and Outfit statistics are summarized for the different item types in Table 3 for the operational items, the pretest variants, and the templates. The expected values of the mean-square Infit and Outfit statistics are both 1.0. Values between .7 and 1.3 are considered, by convention, to be "acceptable." All three item types have mean fit statistics very close to 1.0. However, the operational items have much more variation than either the pretest variants or the templates. The effect size (ratio of standard deviations) when comparing the MS Infit for the Operational items to the MS Infit for the pretest variants is 3.02. The same calculation for MS outfit is 2.47. The effect sizes for the comparison of the operational items to the templates are 4.01 and 4.34 for MS Infit and MS Outfit respectively. Figures 3 and 4 illustrate how much more spread out (along the vertical axis) both the Infit and Outfit are for the operational items.

An important implication of the results in Table 3 and Figures 3 and 4 is that the pretest items, on average even at the individual item/template level, actually fit the Rasch model better than the operational items. This is encouraging insofar as suggesting that the templating process certainly did not add any additional nuisance error or method variance. In fact, the process of

templating may have actually reduced the amount of misfit typically experienced for items on this examination.

Table 3

Mean Square Infit and Mean Square Outfit by Item Type

Descriptive Statistics

Item Status		N	Minimum	Maximum	Mean	Std. Deviation
Pretest Variants	MS Infit	81	.81	1.13	.9977	.06023
	MS Outfit	81	.76	1.34	1.0084	.09922
Operational Items	MS Infit	98	.71	2.11	1.0246	.18218
	MS Outfit	98	.59	2.48	1.0217	.24527
Templates	MS Infit	14	.91	1.05	1.0093	.04548
	MS Outfit	14	.90	1.07	1.0171	.05649

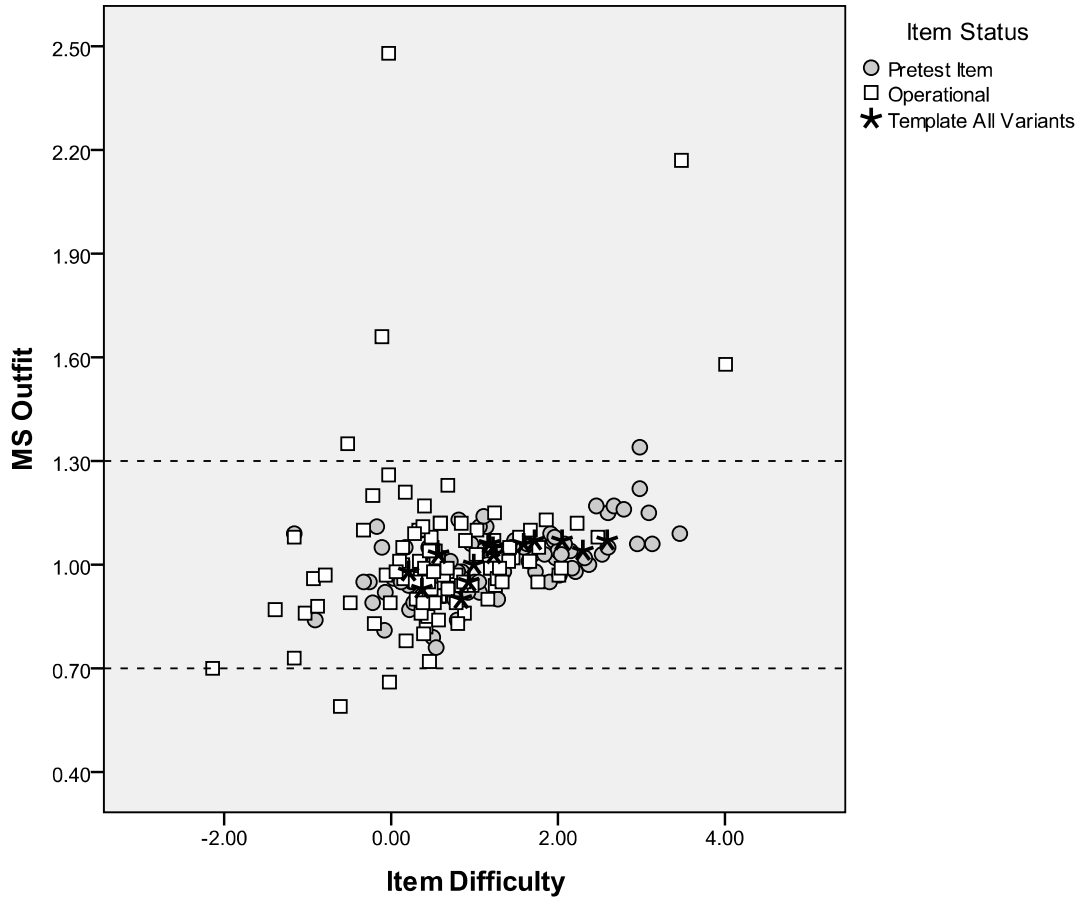


Figure 3. Scatter Plot of Mean Square Outfit Versus Difficulty by Item Type

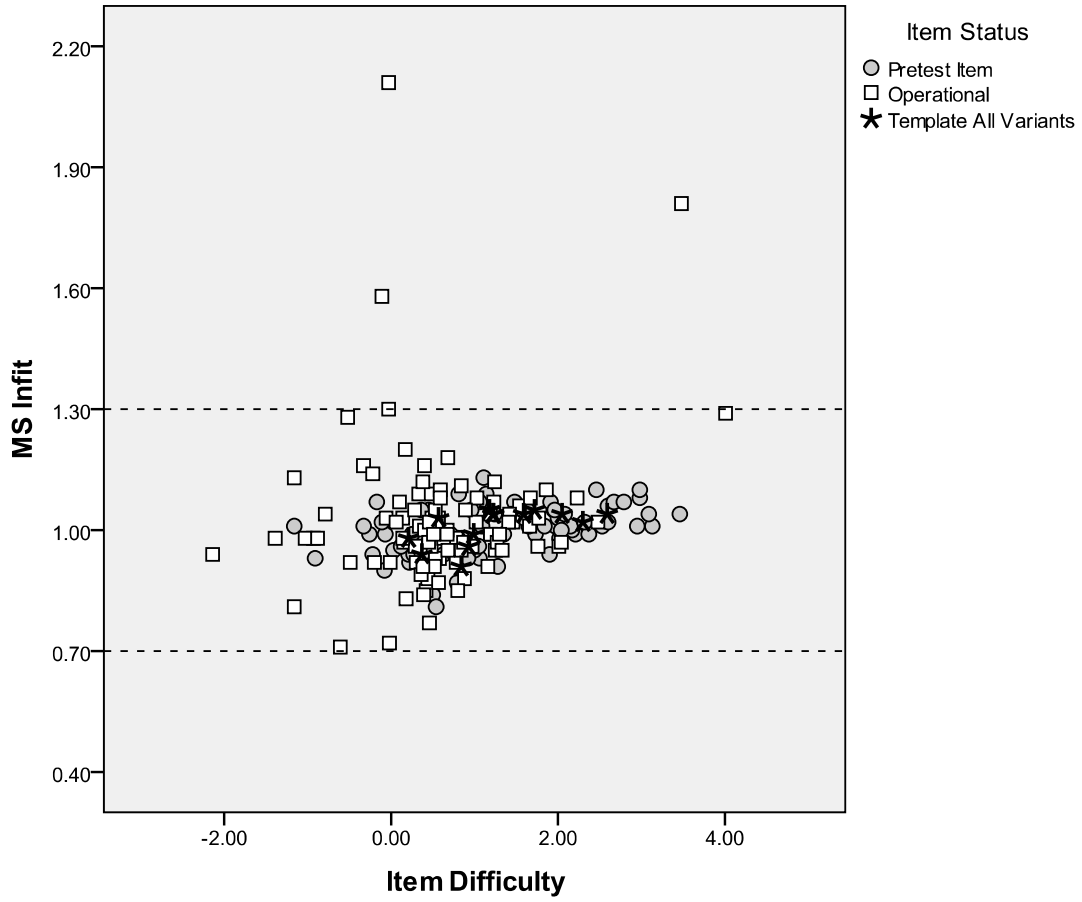


Figure 4. Scatter Plot of Mean Square Infit Versus Difficulty by Item Type

Similarity of Classical and IRT Statistics Within Template

The second research question asked whether the item variants developed from the templates would yield similar classical and item-response theory statistics. Classical statistics included item means (proportion-correct or p-values) and item-total product-moment correlations (i.e., point-biserial correlations for dichotomously scored items). The IRT statistics included Rasch item difficulties (b_i in Equation 4-1), standard error of estimate for those item

difficulties, MS Infit, MS Outfit (described above), and a point-measure product-moment correlation computed by Winsteps[3]. Table 4 shows that the summary statistics for the pretest variants are very similar to the operational items for both the classical and the IRT statistics. The standard errors of estimate for the operational item difficulties are noticeably smaller for the operational items, but that is because the samples for the operational items are, on average, approximately four times as large as the samples for the pretest variants[4]. In practice, for this examination program, any pretest item that has a point-measure correlation less than 0.10 does not normally make it to an operational status unless the item is in an area where the pool is especially weak and the SME committee confirms that the key is unambiguously correct (i.e., that there are no secondary, partially correct keys).

Table 5 shows the summary statistics for the pretest item variants, broken down for each of the 14 templates. Figure 5 graphically illustrates the spread of Rasch item difficulty estimates for each template. Figure 6 graphically portrays the spread of the p-values for each template. The template-level item difficulty estimates are also shown in the plots and, as expected, roughly approximate the central tendency of the distribution of variant-level statistics for each template.

[3] The *point-measure* correlation is a biserial correlation between the responses, u_{ij}, and estimated examinee proficiency scores, θ_j, computed for each item. Note that, on short tests, this biserial correlation may be inflated due to auto-correlation, since the response, u_{ij}, is also used in the computation of the proficiency estimate of θ.

[4] Standard errors of estimate are directly proportional to the sample size (number of examinees). There were far more examinees taking each operational item, which lowered those standard errors.

Table 4

Summary Statistics Across Item Types

Item Status		No. of Items	Minimum	Maximum	Mean	Std. Deviation
Pretest Item	Sample (No. Examinees)	81	104	403	296.26	63.3693
	P-Value	81	.12	.91	.52	.2064
	Point Biserial Correlation	81	-.06	.52	.17	.1165
	Item Difficulty	81	-1.16	3.46	1.23	1.0241
	Standard Error	81	.11	.22	.14	.0258
	MS Infit	81	.81	1.13	1.00	.0602
	MS Outfit	81	.76	1.34	1.01	.0992
	Point Measure Correlation	81	.01	.57	.25	.1108
Operational	Sample (No. Examinees)	98	1075	3287	1308.03	515.4006
	P-Value	98	.10	.97	.65	.1663
	Point Biserial Correlation	98	-.04	.37	.18	.0857
	Item Difficulty	98	-2.14	4.01	.60	.9227
	Standard Error	98	.04	.16	.07	.0172
	MS Infit	98	.71	2.11	1.02	.1822
	MS Outfit	98	.59	2.48	1.02	.2453
	Point Measure Correlation	98	.05	.42	.26	.0810

Another way to evaluate the performance of the pretest variants is to compare the percentage of pretest variants that meet the Pearson VUE criteria for promotion from pretest status to operational status with the historical percentage of pretest items developed in traditional ways that meet the Pearson VUE criteria. There are three criteria that each pretest item must meet: (a) the

PTME must be greater than or equal to 0.10, (b) the P-Value must be greater than 0.25, and (c) the P-Value must be less than 0.95. For the pretest variants, 85.2% meet these criteria and 92.9% of the templates meet the criteria. Historical results show that approximately 75% of items developed in the traditional ways meet the criteria[5].

To evaluate the variance of the pretest variants for each template, it is necessary to apply a correction factor to account for the differences in sample sizes for the pretest variants and for both the operational items of similar difficulty and the templates. The average sample size of the pretest items was a little less than 300, while the average sample size for the operational items was about 1,300. The average sample size for the templates when calibrated using all variants was about 1,700. For each template, three variation values were calculated: SE_T * (sample sized adjusted Standard Error of Template), SE_O * (sample-sized adjusted Standard Error of Operational Items of similar difficulty), and SD_v (standard deviation of item variant difficulty estimates)[6]. To determine the adjustment factors, it was necessary to calculate $\overline{n}_o$, the average sample size of the operational items of similar difficulty (the five nearest to the template difficulty) and $\overline{n}_v$, the average sample of the item variants associated with the template. The three variation values were calculated with the following formulas:

[5] Personal Communication with Pearson VUE psychometricians.
[6] The standard deviation of the difficulty estimates for the variants within a template is an empirical standard error.

$$SE_T^* = SE_T \sqrt{\frac{n_T}{\overline{n}_v}} \tag{4.2}$$

$$SE_O^* = \overline{SE}_O \sqrt{\frac{\overline{n}_o}{\overline{n}_v}} \tag{4.3}$$

$$SD_v = s.d.(variants) / (n-1) \tag{4.4}$$

Figure 7 shows the three variation values for each template plus the average standard error for the variants associated with the templates. Those templates where SD_v is similar in magnitude to the other three values, the templating has worked very well. This is true for many of the templates. However for a few of the templates (1, 8, and 12) SD_v is quite a bit larger than any of the other measures. There are many possible reasons that could explain why a particular item might vary from the other items in a given template. For example, the Illinois Insurance SMEs identified one of the options on the most difficult item from Template 1 as being correct per the regulations, but obtuse in practice, and recommended revising it after the study was complete. The SMEs also identified one option that was used in three of the items from Template 8 that is correct per the regulations, but may not be enforced. This may have led to confusion among the candidates. The SMEs recommended a revision to the key on the easiest item from Template 12, which would have made the item more difficult. In a normal situation, all of these issues would have been fixed prior to the forms

being published. While these issues likely would have been fixed, if they had not, then a graph similar to Figure 7 would be a flag that the item writer(s) assigned to those templates had not been following the rules and more training might be required.

Table 5

Summary of Item-Level Statistics for Each of 14 Templates

Template	N		Minimum	Maximum	Mean	Std. Deviation
I89281	5	Item Difficulty	-.91	1.28	.2400	.91796
		Standard Error	.11	.19	.1420	.03114
		MS Infit	.91	.99	.9420	.03033
		MS Outfit	.84	.96	.9140	.04775
		Point Measure Correlation	.24	.43	.3380	.07694
I89282	8	Item Difficulty	.22	2.21	.7875	.61078
		Standard Error	.12	.22	.1588	.04824
		MS Infit	.81	.99	.9050	.06071
		MS Outfit	.76	.98	.8800	.07982
		Point Measure Correlation	.28	.57	.4175	.10181
89283	8	Item Difficulty	-.08	2.17	.9363	.79383
		Standard Error	.11	.15	.1313	.01246
		MS Infit	.90	1.05	.9750	.05425
		MS Outfit	.81	1.07	.9738	.08618
		Point Measure Correlation	.18	.42	.3000	.08586

Template	N		Minimum	Maximum	Mean	Std. Deviation
I89284	3	Item Difficulty	1.97	2.60	2.3133	.31880
		Standard Error	.12	.14	.1300	.01000
		MS Infit	.99	1.06	1.0200	.03606
		MS Outfit	1.00	1.15	1.0567	.08145
		Point Measure Correlation	.14	.26	.2167	.06658
I89286	4	Item Difficulty	-.22	.45	.2000	.29698
		Standard Error	.12	.14	.1325	.00957
		MS Infit	.94	1.05	.9750	.05196
		MS Outfit	.89	1.10	.9675	.09142
		Point Measure Correlation	.17	.34	.2900	.08042
I89287	8	Item Difficulty	-.07	2.02	.9263	.72744
		Standard Error	.13	.16	.1413	.01126
		MS Infit	.94	.99	.9588	.02167
		MS Outfit	.89	.98	.9338	.03335
		Point Measure Correlation	.25	.38	.3325	.04892
I89288	5	Item Difficulty	-.17	2.15	1.2260	.94849
		Standard Error	.12	.14	.1260	.00894
		MS Infit	.98	1.07	1.0160	.03578
		MS Outfit	.98	1.11	1.0300	.05385
		Point Measure Correlation	.09	.32	.2320	.09094
I89289	5	Item Difficulty	1.48	3.46	2.5560	.88974
		Standard Error	.11	.17	.1460	.02302
		MS Infit	.99	1.08	1.0380	.03834
		MS Outfit	.98	1.22	1.0840	.08678
		Point Measure Correlation	.04	.30	.1520	.09757

Template	N		Minimum	Maximum	Mean	Std. Deviation
18990		Item Difficulty	1.06	2.08	1.5940	.48206
		Standard Error	.12	.12	.1200	.00000
	5	MS Infit	1.00	1.09	1.0420	.03834
		MS Outfit	1.01	1.11	1.0640	.04561
		Point Measure Correlation	.11	.28	.2000	.07583
I89291		Item Difficulty	-.11	1.91	1.1400	.73081
		Standard Error	.11	.19	.1350	.03017
	6	MS Infit	1.01	1.09	1.0433	.03077
		MS Outfit	1.03	1.13	1.0650	.03782
		Point Measure Correlation	.13	.26	.1900	.05254
I89292		Item Difficulty	.25	2.32	1.1000	.79073
		Standard Error	.11	.19	.1340	.03209
	5	MS Infit	.99	1.13	1.0340	.05771
		MS Outfit	.95	1.14	1.0320	.07050
		Point Measure Correlation	.10	.28	.2160	.07162
I89293		Item Difficulty	-1.16	1.96	.3600	1.19442
		Standard Error	.12	.20	.1460	.03435
	5	MS Infit	.94	1.05	1.0120	.04494
		MS Outfit	.91	1.09	1.0180	.08228
		Point Measure Correlation	.12	.37	.2060	.09711
I89294		Item Difficulty	.52	2.79	1.7900	1.00425
		Standard Error	.12	.14	.1267	.01033
	6	MS Infit	.99	1.10	1.0433	.04367
		MS Outfit	1.01	1.17	1.0950	.07994
		Point Measure Correlation	.06	.28	.1633	.10443

Template	N		Minimum	Maximum	Mean	Std. Deviation
I89295	8	Item Difficulty	.12	3.09	2.2175	1.00074
		Standard Error	.11	.19	.1450	.02828
		MS Infit	.96	1.10	1.0250	.04036
		MS Outfit	.95	1.34	1.0838	.11710
		Point Measure Correlation	.01	.30	.1925	.08548

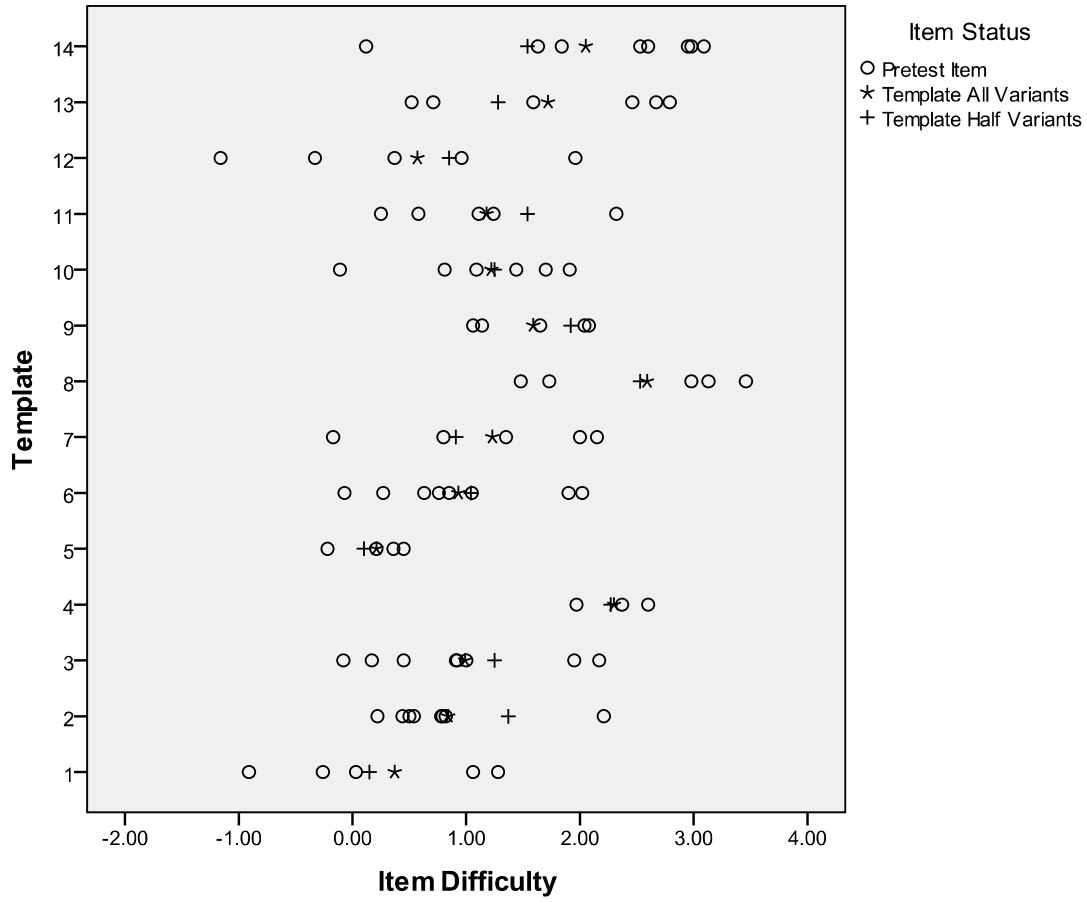


Figure 5. Scatter Plot of Item Difficulty Estimates by Template

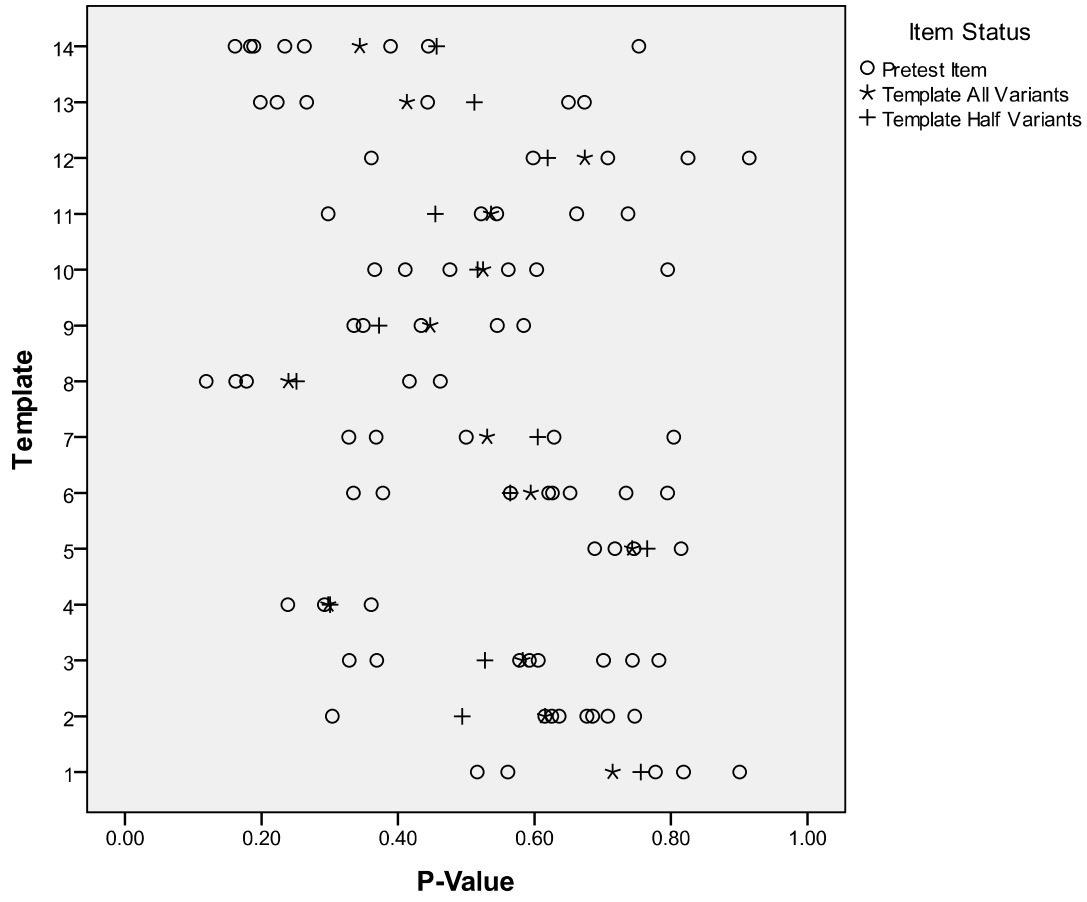


Figure 6. Scatter Plot of Item P-Values by Template

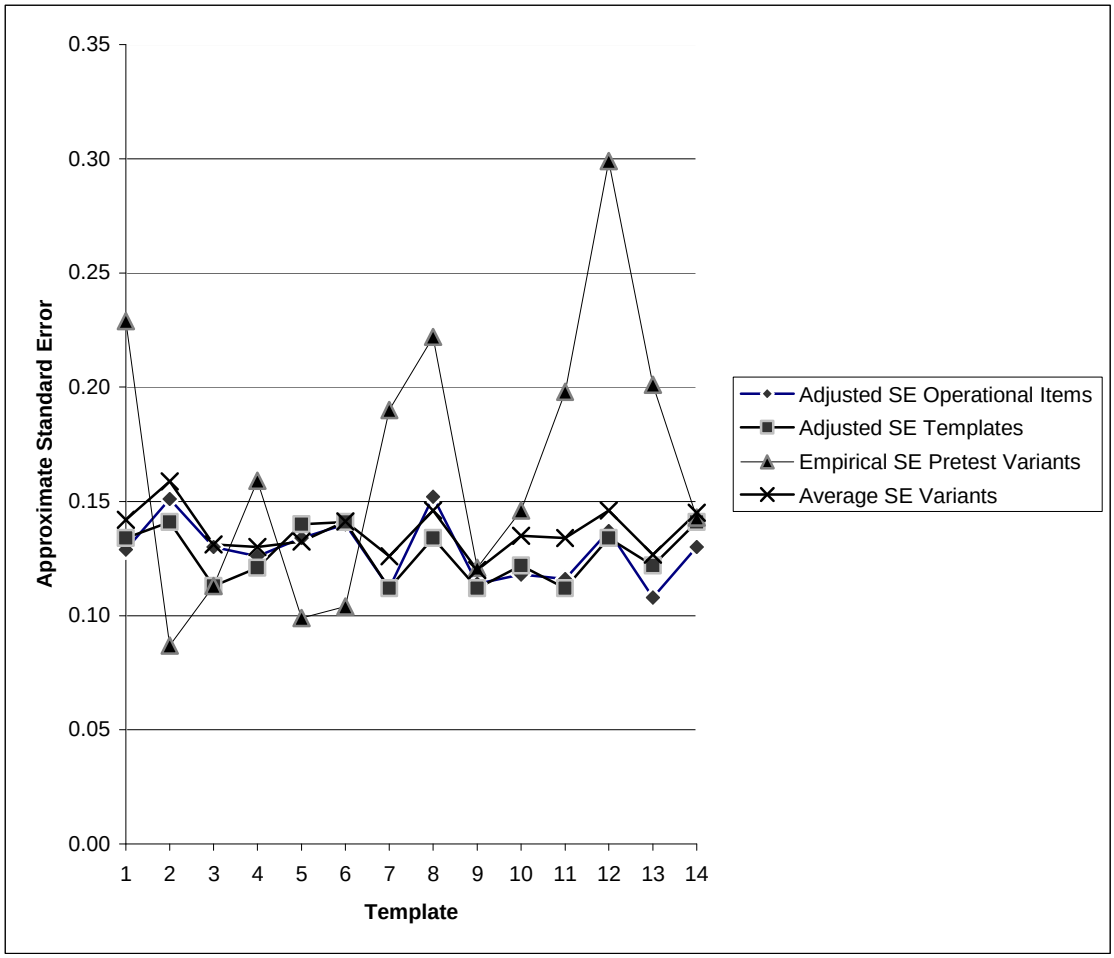


Figure 7. Item Difficulty Variation by Template

Utility of Logical (SME-Determined) Difficulties

The third research question evaluated whether logically-determined item difficulty might be used to replace or at least supplement item difficulty estimates computed using empirical data as well as the effect of using those logically determined difficulty parameters to estimate examinee scores. The logically determined item difficulty parameters (i.e., item complexity scores based on ratings of surface-level features of the items and difficulties based on frequency and criticality ratings from an informal practice analysis of the examination content) were computed in two ways. The first method considered the frequency and importance ratings obtained from the informal practice analysis (see chapter III). The second method considered item complexity scores based on the surface-level complexity of the distractor options, augmented by the positive/negative orientation of the item stem.

One aspect of this research question considered the apparent relationship between the logically determined difficulty parameters (frequency/importance based versus item-complexity rating based) and empirically based item difficulty parameters (i.e., Rasch item difficulties estimated using real data). A second aspect of this research question, conditional on at least showing a reasonable relationship between the logically and empirically based item difficulties, would be to demonstrate the impact of using the former [logically determined] parameters to actually score the examinees. However, a formal analysis of the

second aspect proved to be unnecessary since the logically determined difficulty parameters failed to show any useful relationship to the empirically based difficulty estimates. It was, therefore, not useful to explore the potential use of either set of logically determined difficulty parameter estimates to evaluate potential impact on proficiency scoring.

As described in chapter III, the informal practice–analysis was carried out by having SMEs provide frequency and criticality (importance) ratings for each level of the content blueprint. The SMEs also logically determined the complexity of the items (i.e., provided complexity ratings for each item). The frequency and criticality ratings of the test content levels from the informal practice analysis were correlated with the average Rasch difficulty estimates computed for each corresponding content level (the latter computed using only the operational item difficulty estimates[7]). The correlation between the SME frequency rating and the averages of the operation item difficulties per content level is -.327, suggesting a low, negative correlation between the prevalence of encountering the content levels and average difficulty of those levels. Although the frequency rating is significantly different than zero (at a 0.05 level of significance), the correlation is too low to be practically useful. The correlation between the SME criticality rating and the empirical item difficulty estimates is only .012. Obviously the criticality

[7] A limitation of this study was that the SMEs were only able to rate the frequency and criticality of the content areas, not the individual items or templates. By subsequently needing to average the item difficulties within content level, and because the content levels did not appear to vary substantially in average difficulty, it is not surprising that there was only spurious statistical relationships (correlations) between the content levels and average empirical item difficulties.

ratings do not add to the explained variance for item difficulty. This lack of variation (range restriction) easily explains the low correlations. However, this outcome is somewhat predictable, given the way in which the criticality (and frequency) ratings are typically obtained. Specifically, the SMEs used in this study had a role in developing the content blueprint. It is, therefore, natural that they would rate the importance of most of the examination content as "essential." In retrospect, that finding may seem obvious. However, the informal practice analysis used in this study mimics the types of data that are routinely obtained from a formal practice analysis. This finding confirms that content-based ratings of importance and prevalence may not be very useful in designating surrogate item difficulties.

The complexity scoring scheme did not fare any better insofar as showing a relationship with the empirical item difficulty estimates. The option ratings (i.e., whether the stem of the pretest variant was positive or negative) explained less than 1% of the variance in empirical item difficulty (R^2 unadjusted = 0.007)—not enough to provide any practical utility. This is not to say that SMEs are incapable of generating useful task-model or template estimates of difficulty, only that the relatively simple option rating scheme employed in this study was not the most effective scheme that might have been used. For example, Luecht, Burke, and Devore (2009) demonstrated that a more elaborate cognitive complexity item rating system was very effective in predicting empirical item difficulties. In their

study, however, more complex item types were used which allowed for isolating the key cognitive difficulty factors across different task models. In any case, discovery of more effective ways of rating the cognitive complexity of items may prove useful in future research. This issue of designing complexity into the items as a means of controlling difficulty is discussed in greater depth in chapter V.

Impact of Calibration Strategies on Proficiency Scores

As noted in the introductory section of this chapter, a key aspect of this study relates to any impact and potential utility of using the item templates' difficulty estimates to estimate candidate scores. That is, if each template-level difficulty estimate can be used for all items generated under that template, it is no longer necessary to pretest every item and a natural quality control mechanism of managing variation within template can be implemented[8].

Figures 8 to 12 show bar graphs of the person proficiency estimates computed using four different estimates of each item difficulty: (a) estimates based only on the operational items; (b) estimates based only on the pretest items, separately calibrated as unique items; (c) estimates at the template level; and (d) cross-validation estimates (i.e., calibrating the template difficulty estimates using only an approximately half-sized subset of pretest items and

[8] In practice, strong quality controls for item writing are largely nonexistent. Provided the item writer generates items that match particular content codes and that have minimally acceptable statistics (e.g., positive item–total correlations), the item writer is considered to be doing his or her job. Under this newer assessment engineering paradigm, the item writer's goal is very concrete and measurable: To reduce the variation of item difficulties within each template. A broader range of difficulty is accomplished by devising new templates of more or less complexity.

then using those difficulties to estimate the proficiency for examinees taking the remaining pretest items). With the Rasch IRT model, both the person ability and the item difficulty are on *logit* scale[9]. Therefore, to maximize the reliability of an exam, the test developer should strive to come as close as possible to matching the distribution of item difficulties to the candidate population[10]. This is not as critical for a criterion-referenced exam, but it is still best to separate the candidates as much as possible.

Figure 8 shows the mirrored distributions of proficiency scores and item difficulties for the operational items only. It is clear that, if these items are representative of the complete item bank and if the candidates who tested between July and December 2010 are representative of the candidate population, the bank is not targeted very well to the candidate population. There are too few difficult items that would provide useful information about candidates at abilities above about 1.0 logits. However, as noted in Footnote 4, it is possible to also target the item difficulties in the region of the cut score. In any case, we could develop more difficult item templates that would increase the precision of scores near the upper regions of the proficiency distribution, if that were a desired test specification outcome. Figure 9 shows the impact of adding in the pretest items from the templates. Those additional items do provide some

[9] A logit is defined as the natural logarithm of the odds ratio of right to wrong answers: $\lambda=\ln[P_{ij}/(1-P_{ij})$. Logits typically range from −5 to 5 and simplify the probabilistic interpretation of performance at different points of the scale.

[10] For a mastery test, the concentration of item difficulties can also be effectively targeted near the pass/fail cut point (Luecht, 2006).

apparent improvement in matching items to the examinee proficiencies. However, there are still too few difficulty items.

Figure 10 plots only the item difficulties and examinee proficiency of the pretest items (separate calibration of each item). It comes very close to matching the distribution of the candidate abilities. Figures 11 and 12 show that the 14 item templates can produce items at almost all points along the logit scale, whether the templates are calibrated using all of the item variants (within template) or just a subset of them.

The four calibration strategies produce four corresponding sets of item difficulty estimates (estimates based only on the operational items, estimates based only on the pretest items, separately calibrated as unique items, estimates based on calibrations at the template level using all variants, and estimates based on calibrations at the template level using the cross-validation technique).

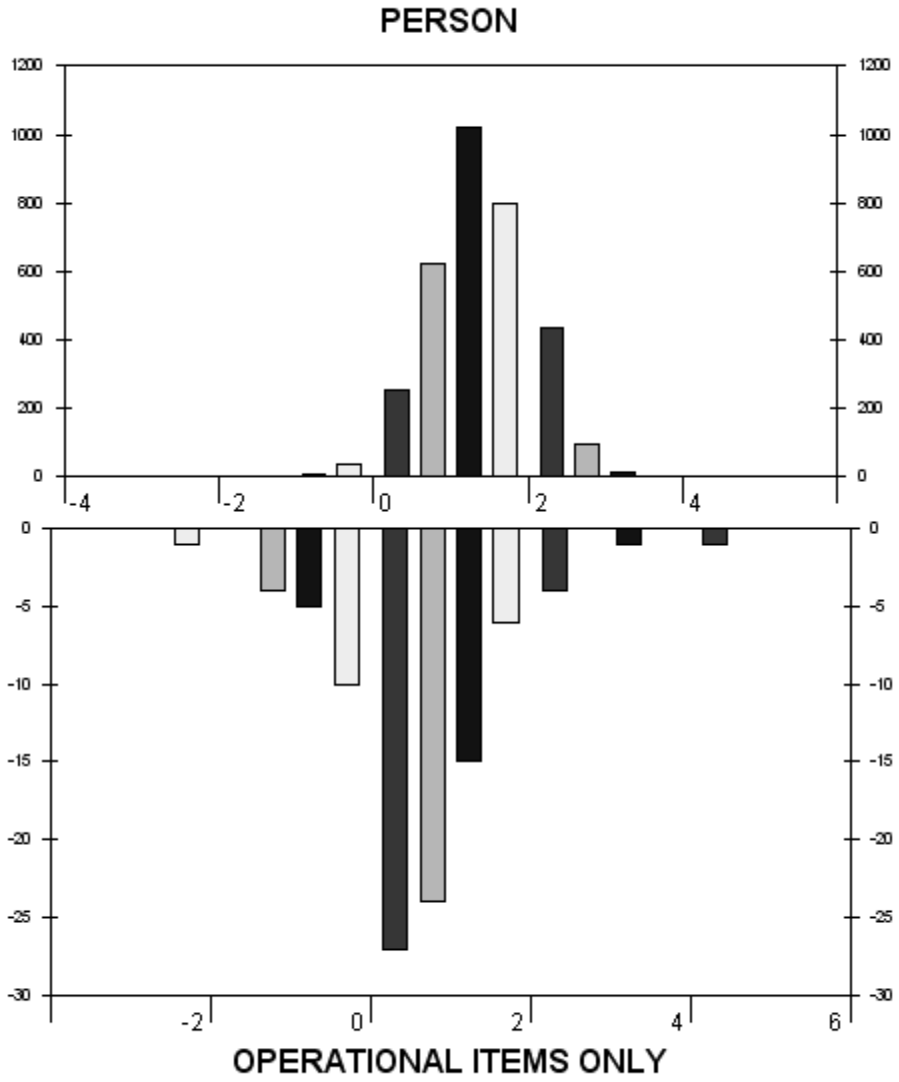


Figure 8. Item-Person Plot – Operational Items Only

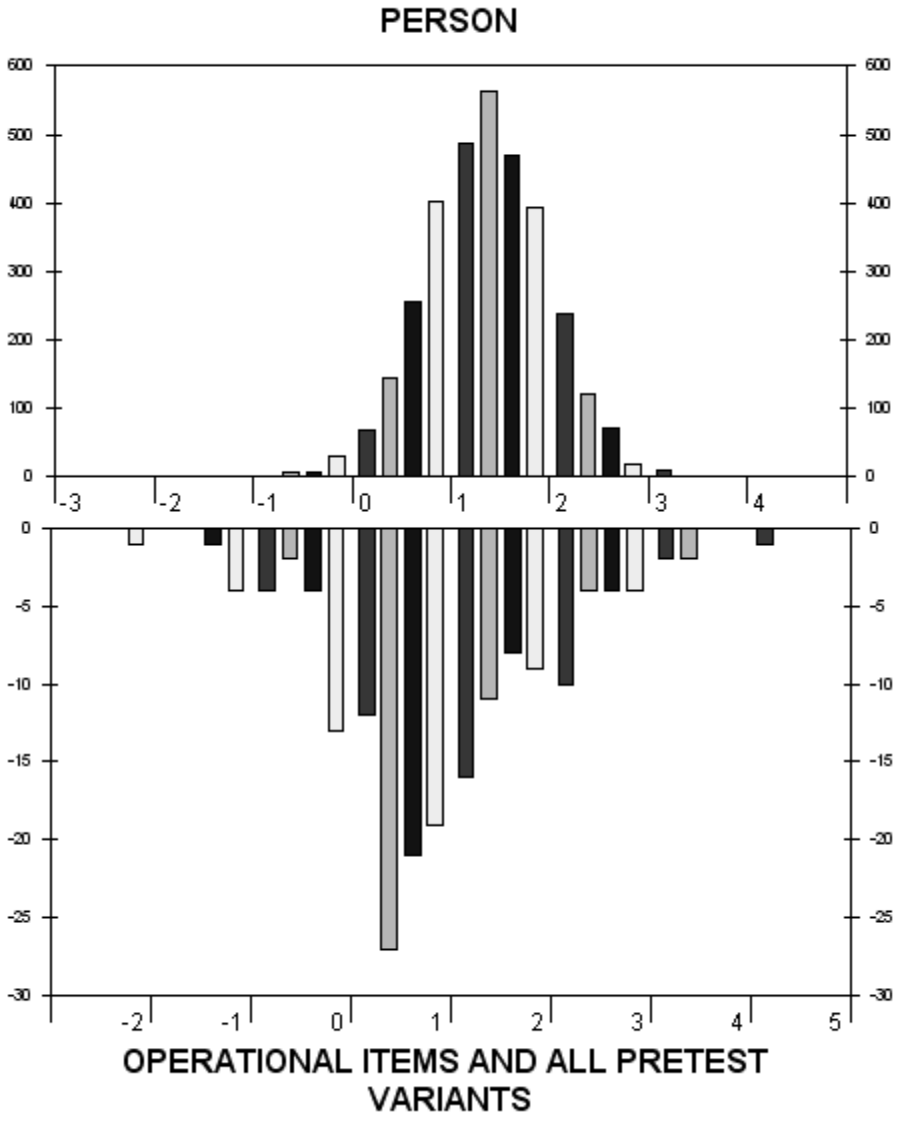


Figure 9. Item-Person Plot – All Operational Items and Variants

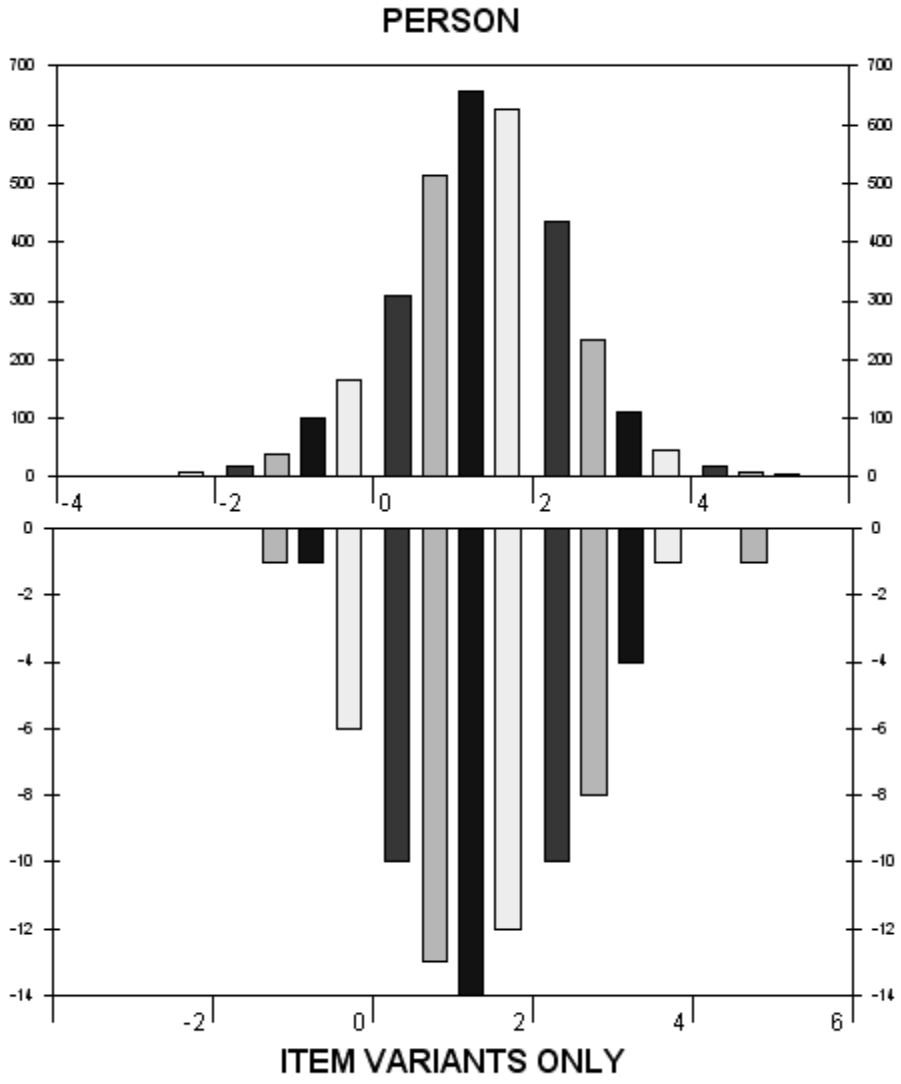


Figure 10. Item-Person Plot–Item Variants Only

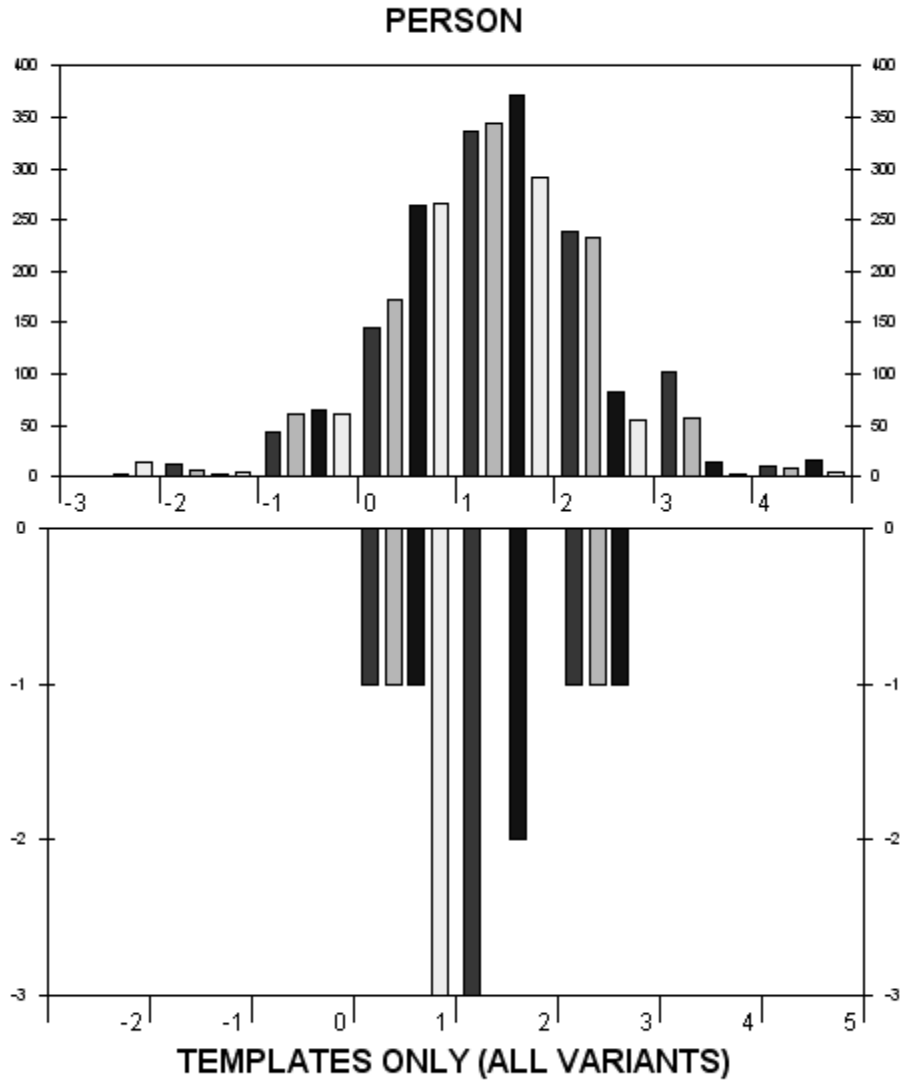


Figure 11. Item-Person Plot – Templates Only (All Variants)

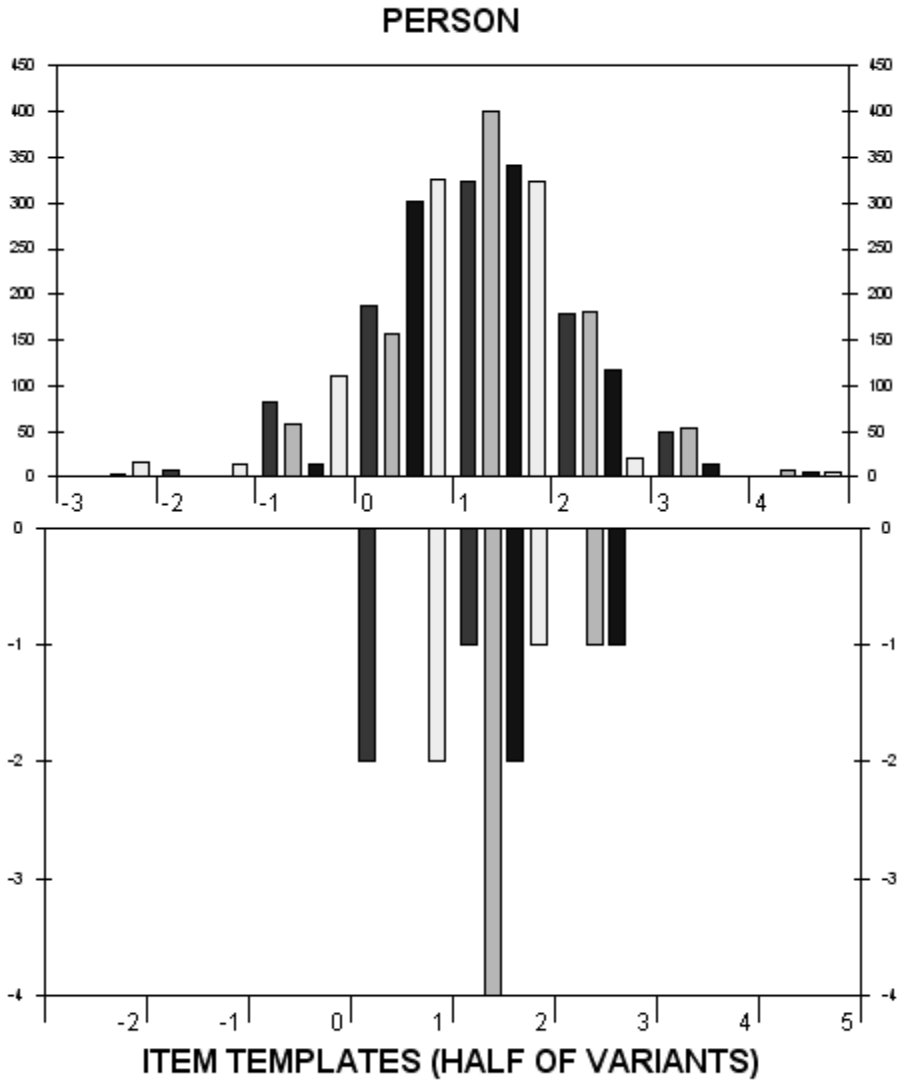


Figure 12. Item-Person Plot – Templates Only (Half of Variants)

Table 6 shows the difficulty ratings and other statistics for each of the 14 templates based on collapsing the data across all of the pretest item variants (calibration strategy number 3). Table 7 shows the same information except that only a random subset of approximately half of the variants was used (calibration strategy number 4).

Table 6

Summary Statistics Across Templates (All Variants)

	Item Difficulty	Standard Error	MS Infit	MS Outfit	Point Measure Correlation
I89281	.37	.06	.94	.93	.34
I89282	.84	.05	.91	.90	.41
I89283	.99	.04	.99	1.00	.28
I89284	2.30	.07	1.02	1.04	.22
I89286	.21	.07	.98	.98	.28
I89287	.93	.05	.96	.95	.33
I89288	1.23	.05	1.04	1.03	.22
I89289	2.59	.06	1.04	1.07	.16
I89290	1.59	.05	1.04	1.06	.21
I89291	1.22	.05	1.04	1.05	.20
I89292	1.18	.05	1.05	1.06	.19
I89293	.57	.06	1.03	1.03	.20
I89294	1.72	.05	1.05	1.07	.18
I89295	2.05	.05	1.04	1.07	.19

Table 7

Summary Statistics Across Templates (Half Variants)

	Item Difficulty	Standard Error	MS Infit	MS Outfit	Point Measure Correlation
I89281	.15	.08	.95	.92	.33
I89282	1.37	.08	.94	.94	.37
I89283	1.25	.06	.97	.96	.33
I89284	2.27	.09	1.04	1.07	.20
I89286	.10	.09	.99	1.02	.24
I89287	1.04	.07	.96	.95	.34
I89288	.91	.07	1.05	1.05	.19
I89289	2.53	.07	1.03	1.06	.19
I89290	1.92	.07	1.02	1.03	.25
I89291	1.25	.07	1.05	1.05	.20
I89292	1.54	.07	1.07	1.08	.16
I89293	.85	.07	1.06	1.07	.16
I89294	1.28	.06	1.03	1.03	.23
I89295	1.54	.06	1.04	1.04	.20

The candidates were actually scored in four different ways: (a) using only the operational items; (b) using only the eight pretest item variants actually administered to each candidate; (c) using the template-calibrated difficulty estimates based on all variants instead of only the pretest variants delivered; and (d) using the cross-validation, template-calibrated item statistics. The

distributions of person ability for these four scoring methods are shown in Figures 6, 8, 9, and 10 respectively.

Table 8 shows the correlations of the ability estimates for each of the four scoring strategies. Note that the three latter scoring strategies involve estimating Rasch proficiency scores based on only eight items. It is, therefore, expected that the estimates will be less reliable than the estimates based on the operational items. It is also expected that lower reliability will attenuate (lower) the correlations between those estimates and any other estimates. It is very encouraging that the two strategies based on using the calibration of the templates correlated slightly better than the strategy based on the individual pretest variants.

Table 8

Correlation Among Different Ability Estimates

	Strategy 1	Strategy 2	Strategy 3	Strategy 4
Strategy 1 - All Operational Items	1	.352	.367	.369
Strategy 2 - Pretest Variants Only	.352	1	.888	.914
Strategy 3 -Templates Only - All Variants	.367	.888	1	.946
Strategy 4 - Templates Only - Half Variants	.369	.914	.946	1

CHAPTER V

CONCLUSIONS AND DISCUSSION

Summary and Implications of Findings

The primary purpose of this thesis is to assess the practicality of using Assessment Engineering to improve the test development processes involved with item development in a licensure context. In this study, existing operational items from a licensure testing program were reverse engineered to develop item templates that could, in turn, be used to develop multiple extended-matching, true-false items in specific areas of the existing content blueprint where the item pool was particularly weak. Using 14 templates, 81 items were developed and field tested for a period of approximately six months; that is, each of the templates was used to create multiple items (variants). The items were then individually pilot tested during an operational administration of the licensing examination. Overall these newly developed, template-based items performed as well or better than the operational items that had been developed in traditional ways. It was also shown that smaller samples of candidates seeing a few items from a template can be used to assess the overall difficulty of the template which, in turn, can be used to target new item development at difficulty levels where measurement information is required, given the distribution of candidates.

The first research question asked whether item variants developed using AE methods would fit a Rasch calibration/scaling model. Using the Winsteps fit statistics (mean square infit and mean square outfit), the results clearly showed that the item variants and the associated templates have less apparent misfit (residual covariance) than the operational items. For example, Figure 3 in chapter IV, shows that only one item variant and no templates had MS outfit values outside the range usually considered acceptable–between 0.7 and 1.3–while there are seven operational items outside that range. Similarly, Figure 4 shows that none of the item variants or templates had MS infit statistics outside that same acceptability range, while there are three operational items that would be classified as somewhat misfitting the Rasch model. These results demonstrate that, for this study, the item variants do indeed fit a Rasch calibration/scaling model quite well and somewhat control misfit (residual covariance). In any case, the templating process applied here does not add any additional nuisance error or method variance that might detract from the quality of the item parameter estimates.

The second research question asked whether the item variants would yield similar classical and IRT statistics to one another when aggregated and analyzed within each of the different templates. The templating appeared to work extremely well. For the most part, the amount of variation in the item variant difficulties—that is, the standard deviations of the classical and IRT-based

item difficulties for each template—were more-or-less on par with sample-size-adjusted standard errors of estimate for operational items of similar difficulty, and with the sample-size-adjusted standard errors of estimate for the overall difficulty of each template. This finding demonstrates both the viability of using the template difficulties to represent an entire class of items, as well as the relative merits of using variation in item statistics, as strong quality control mechanisms for evaluating item writers. Figure 7 in chapter IV demonstrates a credible way to monitor the quality of the SMEs who develop the templates and the items from those templates. There is currently no baseline for what is an acceptable amount of variation of the difficulty among the items from a given template. Over time, bars of acceptance could be developed for a given program.

The final research question had two parts: (a) could the logically developed item difficulties be used to replace or at least supplement item difficulty estimates computed using empirical estimates, and (b) what effect would using templates and template-base difficulty estimates have on candidate ability estimates. The practical advantages and cost savings of being able to use subject-matter experts to provide useful item difficulty ratings, versus empirically pilot testing every new item, are obvious. Unfortunately, the methods used to logically develop pseudo item difficulties, based on subject-matter expert ratings of distractor options and ratings of the frequency and criticality of the template content, proved to be ineffective insofar as providing plausible surrogate item

location parameters that might replace empirically estimated difficulties. Range-restriction and other factors inherent in the sampling of SMEs and their training (instructions) were offered as plausible explanations for these results. As noted in chapter IV, the findings confirm that content-based ratings of frequency, importance, and distractor options, at least in the context of this examination, may not be very useful when attempting to predetermine item difficulties (i.e., locations on a scale).

The second part of the final research question considered the impact of using various calibration strategies to estimate the item difficulties and subsequent use of those difficulty estimates in scoring. In order to examine the effect that templates would have on scoring (i.e., estimating candidates' proficiency), the candidates were scored four different ways. The first calibration and scoring strategy used only the 39 operational items to estimate a proficiency score for each examinee. The second strategy scored each examinee using eight pretest items, where the item difficulties were estimated from a concurrent calibration of the 81 individual pretest variants. This individual pretest item calibration obviously involved a sparse data matrix, reflecting the random assignment of item pretest blocks to examinees. Each item difficulty was, therefore, less stable (i.e., calibrated using a smaller sample size), and the proficiency scores were somewhat unreliably based on only eight item responses. The third and fourth methods used the calibrated IRT item difficulties

estimated from an aggregated data set that collapsed all item pretest variant responses for each of the eight templates. For the third method, the template difficulty estimates were based on a calibration of the entire set of candidate responses associated with any item linked to the given template. For the fourth method, the template difficulties were calibrated on the entire set of candidate responses associated with a subset (sample) of the pretest items linked to the given template. Those difficulty estimates were then used in a type of cross-validation design to score the examinees taking the remaining pretest items associated with each template.

As shown in Table 8 (chapter IV), scoring the candidates with either method using the template-calibrated difficulties (i.e., difficulties estimated using the response data collapsed by template) appears to be at least as good as scoring the candidates with the actual pretest variant item difficulty estimates. Further, while candidate proficiency estimates based on only eight item responses are certainly not very reliable, the results in chapter IV certainly suggest that using empirically estimated, template-based item difficulties for an entire class of items generated from each template, is viable. In fact, if the correlations reported in chapter IV were disattenuated to correct for unreliability (estimation errors in scoring), the various calibration and scoring strategies would be somewhat indistinguishable from the perspective of statistical associations.

Limitations of this Study

There were a number of operational and practical constraints that may have impacted the results. First, it was not possible for the full committee of Illinois subject-matter experts to review the items until after they were published. This meant that some items needed to be dropped from the study, and some items were included that could have been improved, given a full review. In an ideal situation, the committee would thoroughly review the templates, including the rules associated with selecting options, prior to the items being developed.

A second limitation addresses the scope of the study in terms of the number of pilot items and templates used, as well as the examinee sample sizes. The use of motivated, “real” examinees was considered to be an essential aspect of this dissertation. However, like most operational licensing examinations, this testing program has strict policy-based restrictions as to the amount of pilot testing that is possible for a single examinee—especially since the primary purpose of the examination is to make accurate pass/fail decisions. The most pretest items that any examinee could see was eight. Given the overall [expected] sample size of the candidate pool, 14 templates and six to eight pretest item variants of each was considered a maximum in order to get at least several hundred responses for each item. The trade-off between having empirical data from real and motivated examinees and evaluating a larger

collection of items may limit external validity in some sense, but definitely improves the internal validity of the study.

A third limitation is that this study used only the extended-matching true-false (EMTF) item type, which requires a list of true options and another list of false options. This item type limits the cognitive complexity of different templates to principally manipulating a small number of declarative knowledge components (e.g., familiarity of terms, negation clauses, and implicit links/cuing). Using an existing, rather simplistic item type like the EMTF, therefore, limited the capability to *systematically* vary the difficulty (complexity) of each template. In fact, strong cognitive models were not developed for the templates—that is a research for the future. Instead, a rather ineffective method (in hindsight) was used to estimate the complexity of each template for EMTF items. Two SMEs rated the option shells for each template on a 5-point scale. However, the SMEs were not trained in developing a common, cognitive definition of item difficulty or complexity. As a result, the rating process was not successful. A task-model-based cognitive method based on strong theory would likely have been much more successful (e.g., Luecht, Burke & Devore, 2009; Leighton & Gierl, 2007;). If a strong cognitively based method used to determine item complexity (and indirectly, template location relative to other templates) had been developed, it might not have been necessary to try to predict the template difficulty based on an informal

practice analysis. That latter method of determining template difficulty was found to be ineffective for this study.

A fourth limitation of the study was the use of only a single template per task model. This limitation would be solved by developing explicit task models in response to the construct maps, rather than reverse-engineering existing items within the context of an existing operational program with a rather explicit content blueprint. Under Assessment Engineering, a new blueprinting process would ideally emerge where content (i.e., cognitive complexity that incorporates content into the task-model specification) and statistical properties of the items (location and sensitivity to the intended proficiency trait) are interchangeable. In addition, as Luecht et al. (2009) noted, “The use of multiple templates makes it possible to have many different *views of the same task model* and greatly expands the possible number of items that can be developed for each task model” (p. 3).

Future Studies

Future studies could be extended to other types of certification and licensure tests. It will also be important to generalize to other traditional and innovative item types, along with the use of strong theory, to develop useful cognitive task models and templates.

REFERENCES

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

Ariel, A., van der Linden, W. J., & Veldkamp, B. P. (2006). A strategy for optimizing item-pool management. *Journal of Educational Measurement*, *43*(2), 85-96.

Ascalon, M. E., Meyers, L. S., Davis, B. W., & Smits, N. (2007). Distractor similarity and item-stem structure: Effects on item difficulty. *Applied Measurement in Education*, *20*(2), 153-170.

Attali, Y., & Bar-Hillel, M. (2003). Guess where: The position of correct answers in multiple-choice test items as a psychometric variable. *Journal of Educational Measurement*, *40*(2), 109-128.

Becker, K., Masters, J. S., & Bailey, J. (2007, April). *Practice effects and program features in professional regulatory examinations*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Bejar, I. I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement*, *7*(3), 303-310.

Bejar, I. I. (1990). A generative analysis of a three-dimensional spatial task. *Applied Psychological Measurement*, *14*(3), 237-245.

Bejar, I. I. (1993). A generative approach to psychological and educational measurement. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 323-358). Hillsdale, NJ: Lawrence Erlbaum.

Bejar, I. I. (1996). *Generative response modeling: Leveraging the computer as a test delivery medium* (ETS Research Report No. 96-13). Princeton, NJ: Educational Testing Service.

Bejar, I. I. (2002). Generative testing: From conception to implementation. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 129-157). Mahwah, NJ: Lawrence Erlbaum.

Bejar, I. I. (2008, February). *Historical perspective*. Session paper at Assessment Engineering: Moving from Theory to Practice. Coordinated panel presentation at the Annual Meeting of the Association of Test Publishers, Dallas, TX.

Bejar, I. I., Braun, H. I., & Tannenbaum, R. J. (2007). A prosecitve, predictive, and progressive approach to standard setting. In R. W. Lissitz (Ed.), *Assessing and modeling cognitive development in school: Intellectual growth and standard setting* (chap. 1). Maple Grove, MN: JAM Press.

Bejar, I. I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2003). A feasibility study of on-the-fly item generation in adaptive testing. *Journal of Technology, Learning, and Assessment*, *2*(3). Available from http://www.jtla.org

Bejar, I. I., & Yocom, P. (1991). A generative approach to the modeling of isomorphic hidden-figure items. *Applied Psychological Measurement*, *15*(2), 129-137.

Belov, D. I., & Knezevich, L. (2008, March). *Predicting item difficulty with semantic similarity measures.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, NY.

Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives: The classification of education goals (Handbook 1: Cognitive domain).* New York, NY: Wiley.

Blumberg, P. (1981/1982). A practical methodology for developing content parallel multiple-choice tests. *Journal of Experimental Education*, *50*(2), 56-63.

Bormuth, J. R. (1970). *On the theory of achievement test items*, Chicago, IL: University of Chicago Press.

Cantor, J. A. (1987). Developing multiple-choice test items. *Training and Development Journal, 41*(5), 85-88.

Case, S. M., & Swanson, D. B. (1993). Extended matching items: A practical alternative to free response questions. *Teaching and Learning in Medicine*, *5*(2), 107-115.

Case, S. M., Swanson, D. B., & Ripkey, D. R. (1994). Comparison of item in five-option and extended-matching formats for assessment of diagnostic skills. *Academic Medicine*, *69*(10), S1-S3.

Chalifour, C. L., & Powers, D. E. (1989). The relationship of content characteristics of GRE and analytical reasoning items to their difficulties and discriminations. *Journal of Educational Measurement*, *26*(2), 120-132.

Cizek, G. J., Webb, L. C., & Kalohn, J. C. (1995). The use of cognitive taxonomies in licensure and certification test development: Reasonable or customary? *Evaluation & the Health Professions*, *18*(1), 77-91.

Clauser, B. E., & Margolis, M. J. (2006). Book Reviews. In S. H. Irvine & P. C. Kyllonen (Eds.) *Item Generation for Test Development*, *International Journal of Testing*, *6*(3), 301-304.

Clauser, B. E., Margolis, M. J., & Case, S. M. (2006). Testing for licensure and certification in the professions. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 701-731). Washington, DC: American Council on Education.

Cromley, J. G., & Mislevy, R. J. (2005). *Task templates based on misconception research.* (PADI Tech. Report No. 6). Menlo Park, CA: SRI International.

Cronback, L. J. (1970). On the theory of achievement test items [Review]. In J. R. Bormuth (Ed.). *Psychometricka*, *35*(4), 509-511.

Davey, T., & Nering, M. (2002). Controlling item exposure and maintaining item security. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 165-191). Mahwah, NJ: Lawrence Erlbaum.

Dennis, I., Handley, S., Bradon, P., Evans, J., &. Newstead, S. (2002). Using the psychology of reasoning to predict the difficulty of analytical reasoning problems. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 129-157). Mahwah, NJ: Lawrence Erlbaum.

Downing, S. M. (2002). Construct-irrelevant variance and flawed test questions: Do multiple-choice item-writing principles make any difference? *Academic Medicine*, *77*(10), S103-S104.

Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education*, *10*(2), 133-143.

Downing, S. M. (2006.) Selected-response item formats in test development. In S. M. Downing, & T. M. Haladyna (Eds.) *Handbook of test development.* Mahwah, NJ: Lawrence Erlbaum.

Downing, S. M., Baranowski, R. A., Grosso, L. J., & Norcini, J. J. (1995). Item type and cognitive ability measured: The validity evidence for multiple-true-false items in medical specialty certification. *Applied Measurement in Education*, *8*(2), 187-197.

Downing, S. M., & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education, 10*(1), 61-82.

Drasgow, F., Luecht, R. M., & Bennett, R. E. (2006). Technology and testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 471-515). Washington, DC: American Council on Education.

Dudycha, A. L., & Carpenter, J. B. (1973). Effects of item format on item discrimination and difficulty. *Journal of Applied Psychology*, *58*(1), 116-121.

Ebel, R. L. (1962). Content standard test scores. *Educational and Psychological Measurement*, *22*(1), 15-25.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, *3*(3), 380-396.

Embretson, S. E. (1999). Generating Items during testing: Psychometric issues and models. *Psychometrika*, *64*(4), 407-433.

Embretson, S. E. (2002). Generating abstract reasoning items with cognitive theory. In S. H. Irvine & P. C. Kyllonen (Eds.) *Item generation for test development*. Mahwah, NJ: Lawrence Erlbaum.

Enright, M. K., & Sheehan, K. M. (2002). Modeling the difficulty of quantitative reasoning items: Implications for item generation. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 129-157). Mahwah, NJ: Lawrence Erlbaum.

Frase, L. T., Almond, R. G., Burstein, J., Kukich, K., Sheehan, K. M., Steinberg, L. S., . . . Chodorow, M. (2003). Technology and assessment. In H. F. O'Neil, Jr. & R. S. Perez (Eds.), *Technology Applications in Education.* Mahwah, NJ: Lawrence Erlbaum.

Frey, B. B., Petersen, S., Edwards, L. M., Petrotti, T., & Peyton, V. (2005). Item-writing rules: Collective wisdom. *Teaching and Teacher Education*, *21*(4), 357-364.

Frisbie, D. A. (1992). The multiple true-false item format: A status review. *Educational Measurement: Issues and Practice*, *11*(4), 21-26.

Frisbie, D. A., & Sweeney, D. C. (1982). The relative merits of multiple true-false achievement tests. *Journal of Educational Measurement, 19*(1), 29-35.

Gierl, M. J., Zhou, J., & Alves, C. (2008). Developing a taxonomy of item model types to promote assessment engineering. *The Journal of Technology, Learning, and Assessment, 7*(2), Retrieved from http://www.jtla.org

Glas, C. A. W. (2006). *Alternative approaches to updating parameter estimates in tests with item cloning.* (Computerized Testing Report No. 03-01). Newtown, PA: Law School Admission Council.

Glas, C. A. W., & van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement*, *27*(4), 247-261.

Gorin, J. S. (2006). Test design with cognition in mind. *Educational Measurement: Issues and Practice, 25*(4), 21-35.

Gorin, J. S., & Embretson, S. E. (2006). Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement*, *30*(5), 394-411.

Graf, E. A. (2008). Approaches to the design of diagnostic item models, (ETS RR-08-07). Princeton, NJ: Educational Testing Service.

Haladyna, T. M. (1992). The effectiveness of several multiple-choice item formats. *Applied Measurement in Education*, *5*(1), 73-88.

Haladyna, T. M. (1994). A research agenda for licensing and certification testing validation studies. *Evaluation & The Health Professions*, *17*(2), 242-256.

Haladyna, T. M. (2002). *Essentials of standardized achievement testing*. Boston, MA: Allyn and Bacon.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed). Mahwah, NJ: Lawrence Erlbaum.

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

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, *15*(3), 309-334.

Haladyna, T., & Shindoll, R. (1989). Item shells: A method for writing effective multiple-choice test items. *Evaluation and the Health Professions*, *12*(1), 97-106.

Hambleton, R. K., & Jirka, S. J. (2006). Anchor-based methods for judgmentally estimating item statistics. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 399-420), Mahwah, NJ: Lawrence Erlbaum.

Hambleton, R. K., Sireci, S. G., Swaminathan, H., Xing, D., & Rizavi, S. (2003). *Anchor-based methods for judgmentally estimating item difficulty parameters.* (Computerized Testing Report No. 98-03). Newtown, PA: Law School Admission Council.

Hamel, L., & Schank, P. (2006). *A wizard for PADI assessment design*. (PADI Technical Report No. 11). Menlo Park, CA: SRI International.

Haynie, K. C., Haertel, G. D., Lash, A. A., Quellmalz, E. S., & DeBarger, A. H. (2006). *Reverse engineering the NAEP floating pencil task using the PADI design system.* (PADI Technical Report No. 16). Menlo Park, CA: SRI International.

Hively, H., Patterson, H. L., & Page, S. H. (1968). A “universe-defined” system of arithmetic achievement tests. *Journal of Educational Measurement*, *5*(4), 275-290.

Hornke, L. F. (2002). Item-generation models for higher order cognitive functions. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development.* Mahwah, NJ: Lawrence Erlbaum.

Huff, K. (2008, February). *The role of EVIDENCE in evidence centered design*. Session paper at Assessment Engineering: Moving from Theory to Practice, Coordinated panel presentation at the Annual Meeting of the Association of Test Publishers, Dallas, TX.

Impara, J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement*, *35*(1), 69-81.

Irvine, S. H. (2002). The foundation of item generation for mass testing. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 129-157). Mahwah, NJ: Lawrence Erlbaum.

Kennedy, C. A. (2008, February). *Building blocks & principles for classroom assessment.* Session paper at Assessment Engineering: Moving from Theory to Practice, Coordinated panel presentation at the Annual Meeting of the Association of Test Publishers, Dallas, TX.

Kyllonen, P. C. (2002). Item generation for repeated testing of human performance. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 129-157). Mahwah, NJ: Lawrence Erlbaum.

LaDuca, A. (1994). Validation of professional licensure examinations. *Evaluation & the Health Professions*, *17*(2), 178-197.

LaDuca, A., Staples, W. I., Templeton, B., & Holzman, G. B. (1986). Item modeling procedures for constructing content-equivalent multiple choice questions. *Medical Education, 20, 53-56.*

Lai, H., Alves, C., & Gierl, M., (2009, June) *Applying item model taxonomy for automatic item generation for CAT*. Poster presented at the 2009 Graduate Management Academic Council Conference for Computerized Adaptive Testing, Minneapolis, MN.

Leighton, J. P. & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice, 26, 3-16.*

Lewis, C. (2002). Discussant remarks. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 129-157). Mahwah, NJ: Lawrence Erlbaum.

Linacre, J. M. (2009). *Winsteps Rasch*-model computer program (Program manual 3.69.0). Downloaded from www.winsteps.com

Luecht, R. M. (1995). *Reorganizing our thinking about test specifications and item content.* Philadelphia, PA: National Board of Medical Examiners.

Luecht, R. M. (2002, April). *From design to delivery: Engineering the mass production of complex performance assessments*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Luecht, R. M. (2003). Multistage complexity in language proficiency assessment: A framework for aligning theoretical perspectives, test development, and psychometrics. *Foreign Language Annals, 36*(4), 518-526.

Luecht, R. M. (2006a) Designing tests for pass-fail decisions using item response theory. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum.

Luecht, R. M. (2006b, April). *Engineering the test: Principled item design to automated test assembly*. Invited special event presentation at the Annual Meeting of the Society for Industrial and Organizational Psychology.

Luecht, R. M. (2007a, April). *Assessment engineering in language testing: From data models and templates to psychometrics*. Invited paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Luecht, R. M. (2007b). Using information from multiple-choice distractors to enhance cognitive-diagnostic score reporting. In J. P. Leighton & M. J. Gierl (Eds), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 319-340). London, England: Cambridge University Press.

Luecht, R. M. (2007c, October). *Assessment engineering: An integrated approach to test design, development, assembly, and scoring.* Invited keynote and workshop presented at the Performance Testing Council Summit, Scottsdale, AZ.

Luecht, R. M. (2007d, June). *Emerging topics (in CBT and CAT)*. Invited keynote presentation at the 2007 GMAC Conference on Computerized Adaptive Testing, Minneapolis, MN.

Luecht, R. M. (2008a, February). *Assessment engineering*. Session paper at Assessment Engineering: Moving from Theory to Practice, Coordinated panel presentation at the Annual Meeting of the Association of Test Publishers, Dallas, TX.

Luecht, R. M. (2008b, February). *The application of assessment engineering to an operational licensure testing program*. Paper presented at the Annual Meeting of the Association of Test Publishers, Dallas, TX.

Luecht, R. M. (2009, June). *Adaptive computer-based tasks under an assessment engineering paradigm*. Paper presented at the 2009 GMAC CAT Conference, Minneapolis, MN.

Luecht, R. M., Burke, M., & Devore, R. (2009, April). *Task modeling of complex computer-based performance exercises*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Luecht, R. M., Gierl, M, Tan, X., & Huff, K. (2006, April). *Scalability and the development of useful diagnostic scales*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Martinez, R. J., Moreno, R., Martin, I., & Trigo, M. E. (2009). Evaluation of five guidelines for option development in multiple-choice item writing. *Psicothema*, *21*(2), 326-330.

Meisner, R., Luecht, R., & Reckase, M. (1993). The comparability of the statistical characteristics of test items generated by computer algorithms. *ACT Research Report Series 93-9*.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13-23.

Millman, J., & Westman, R. S. (1989). Computer-assisted writing of achievement test items: Toward a future technology. *Journal of Educational Measurement, 26*(2), 177-190.

Mislevy, R. J. (1988). Exploiting auxiliary information about items in the estimation of Rasch item difficulty parameters. *Applied Psychological Measurement, 12(3), 281-296.*

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 257-306). Washington, DC: American Council on Education.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centered design* (ETS RR-03-16). Princeton, NJ: Educational Testing Service.

Mislevy, R. J., & Haertel, G. D. (2007, April). *Implications of evidence-centered design for education testing: Lessons from the PADE project*. Invited paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Mislevy, R. J., & Riconscente, M. M. (2005). *Evidence-centered assessment design: Layers, structures, and terminology*. (PADI Technical Report No. 9). Menlo Park, CA: SRI International.

Mislevy, R. J., Sheehan, K. M., & Wingersky, M. (1993). How to equate tests with little or no data. *Journal of Educational Measurement*, *30*(1), 55-78.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (1999). Evidence-centered design. *Educational Testing Service*. Retrieved from http://www.education.umd.edu/EDMS/mislevey/papers/ECDoverview.html

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). On the roles of task model variables in assessment design. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 129-157). Mahwah, NJ: Lawrence Erlbaum.

Newstead, S., Bradon, P., Handley, S., Evans, J., & Dennis, I. (2002). Using the psychology of reasoning to predict the difficulty of analytical reasoning problems. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 129-157). Mahwah, NJ: Lawrence Erlbaum Associates.

Osterlind, S. J. (1998). *Constructing test items: Multiple-choice, constructed-response, and other formats (*2nd ed.*)*, Boston, MA: Kluwer Academic.

Raymond, M. R., & Neustel, S (2006). Determining the content of credentialing examinations. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests. Chicago, IL: University of Chicago Press. (Original work published 1960).*

Richards, J. M. (1967). Can computers write college admissions tests? *Journal of Applied Psychology*, *51*(3), 211-215.

Roid, G. H. (1984, August). *New technologies in the writing of test items*. Paper presented at the Annual Meeting of the American Psychological Association, Toronto, Canada.

Roid, G. H., & Haladyna, T. M. (1982). A technology for test-item writing. Maryland Heights, MO: Academic Press.

Rupp, A. A., Garcia, P., & Jamieson, J. (2001). Combining multiple regression and CART to understand difficulty in second language reading and listening comprehension test items. *International Journal of Testing*, *1*(3&4), 185-216.

Ryan, J. J. (1968). Teacher judgments of test item properties. *Journal of Educational Measurement*, *5*(4), 301-306.

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307-353). Washington, DC: American Council on Education.

Schmitt, K. (1995). What is licensure? In J. C. Impara (Ed.), *Licensure testing: Purposes, procedures, and practices*. Lincoln: University of Nebraska, Buros Institute of Mental Measurements.

Shimberg, B. (1981). Testing for licensure and certification. *American Psychologist*, *36*, 1138-1146.

Shimberg, B., & Roederer, D. (1994). *Questions a legislator should ask.* K. Schmitt (Ed.). Lexington, KY: The Council of State Governments.

Singley, M. K., & Bennett, R. E. (2002). Item generation and beyond: Applications of schema theory to mathematics assessment. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 129-157). Mahwah, NJ: Lawrence Erlbaum.

Spray, J. A., & Huang, C. (2000). Obtaining test blueprint weights from job analysis surveys. *Journal of Educational Measurement*, *37*(3), 187-201.

Swanson, L. (2002). Discussant remarks. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 129-157). Mahwah, NJ: Lawrence Erlbaum.

Swygert, K. A., Scrams, D. J., Thompson, L. E., & Kerman, D. E. (2006). *An evaluation of the impact of cloning on item parameters* (Computerized Testing Report No. 99-08). Newtown, PA: Law *School* Admission Council.

Tarrant, M., & Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical Education*, *42*(2), 198-206.

Wainer, H. (2002). On the automatic generation of test items: Some whens, whys, and hows. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 129-157). Mahwah, NJ: Lawrence Erlbaum.

Wilson, M. (2005). Subscales and summary scales: Issues in health-related outcomes. In J. Libscomb, C. C. Gotay, & C. Snyder (Eds.), *Outcomes assessment in cancer: Measures, methods, and applications* (pp. 465-479), Cambridge, England: Cambridge University Press.

Wright, B., & Stone, M. (1979). *Best test design.* Chicagol, IL: MESA Press.

Wright, B., & Stone, M. (1999). *Measurement essentials* (2nd ed.). Wilmington, DE: Wide Range.

Wright, D. (2002). Scoring tests when items have been generated. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 129-157). Mahwah, NJ: Lawrence Erlbaum.

APPENDIX A. EXEMPLAR TEMPLATE AND ITEMS

Item ID: I89282a-h
Domain: 01020102
Construct Identifier: Knowledge of Rules and Regs

Stem: Limited lines producers are (**NOT**) authorized to sell which of the following products?

	Option Rating*		
	IL SME	PV SME	Overall
TRUE			
1. Baggage Insurance	1	5	3
2. Trip Cancellation Insurance	1	5	3
3. Pre-Paid Legal Services Insurance	1	3	2
4. Limited Travel Health Insurance	3	1	2
5. County Mutual Insurance	3	3	3
6. Industrial Life Insurance	1	5	3
7. Industrial Accident And Health Insurance	1	3	2
8. Legal Expenses Insurance	1	5	3
FALSE			
10. Variable Life Insurance	1	5	3
11. Whole Life Insurance	1	1	1
12. Annuities	1	3	2
13. Pension Plans	5	1	3
14. Group Life Insurance	1	5	3
15. Individual Life Insurance	3	1	2
16. Group Health Insurance	3	1	2
17. Individual Health Insurance	3	1	2
18. Medicare Supplement Policies	3	3	3
19. Term Life Insurance	3	3	3
20. Universal Life Insurance	1	3	2
21. Cancer Insurance	1	1	1
22. Hospital Indemnity Insurance	1	3	2
23. Auto Insurance	1	3	2
24. Group Property And Casualty Insurance	1	5	3

Each item from the template includes a Good key, one Good distracter, and 2 Reasonable distracters. Key is always D.

* 1 Easy, 3 Moderate, 5 Hard

a) (Pos) 10, 16, 12, **1** (Sum=3+2+2+3=10) (Enemies: I08288, I15620, I22426, I22429)
b) (Pos) 22, 17, 13, **5** (Sum=2+2+3+3=10) (Enemies: I06681, I22426)
c) (Pos) 18, 15, 22, **6** (Sum=3+2+2+3=10) (Enemies: I06681)
d) (Pos) 19, 20, 16, **8** (Sum=3+2+2+3=10) (Enemies: I06681, I22429)
e) (Pos) 10, 20, 23, **2** (Sum=3+2+2+3=10) (Enemies: I08124, I22426)
f) (Neg) 2, 3, 7, **24** (Sum=3+2+2+3=10) **(**Enemies: I08288, I08124, I15620)
g) (Neg) 8, 3, 7, **18** (Sum=3+2+2+3=10) **(**Enemies: I08288, I08124, I15620, I22429)
h) (Neg) 1, 3, 7, **19** (Sum=3+2+2+3=10) **(**Enemies: I08288, I08124, I15620, I06681, I22429, I22426)

I89282a

Limited lines producers are authorized to sell which of the following products?

A) Variable Life Insurance
B) Group Health Insurance
C) Annuities
D) Baggage Insurance

I89282e

Limited lines producers are authorized to sell which of the following products?

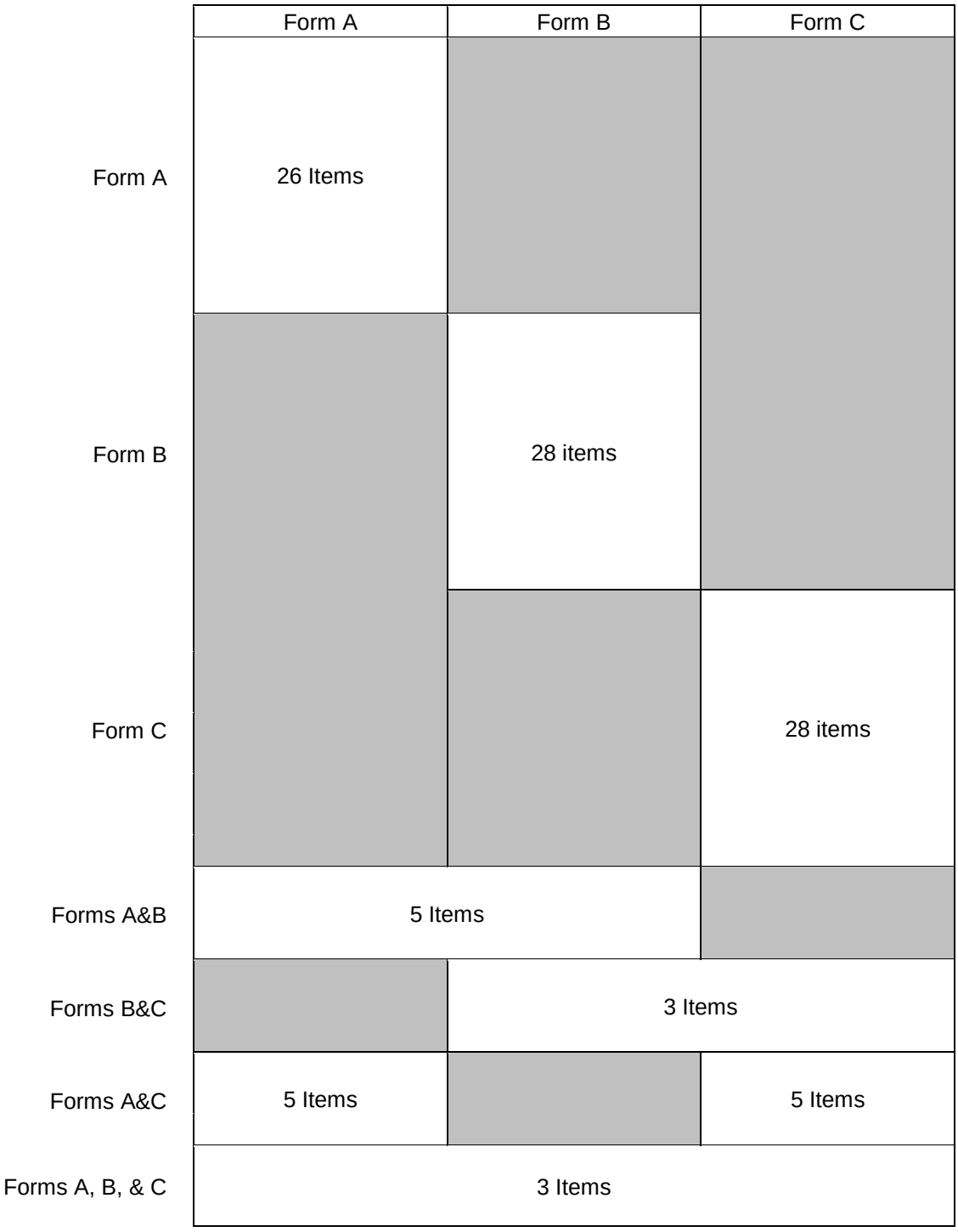
A) Variable Life Insurance
B) Universal Life Insurance
C) Auto Insurance
D) Trip Cancellation Insurance

I89282f

Limited lines producers are **NOT** authorized to sell which of the following products?

A) Trip Cancellation Insurance
B) Pre-Paid Legal Services Insurance
C) Industrial Accident And Health Insurance
D) Group Property And Casualty Insurance

APPENDIX B. OPERATIONAL ITEMS BY FORM



APPENDIX C. PRETEST VARIANTS BY FORM

Template	Form A	Form B	Form C
1	5		5
2	4	1	8
3	6	8	2
4		3	3
5	3	2	3
6	8	8	
7	5		5
8	2	5	2
9		5	5
10	3	5	3
11	4		5
12		5	5
13	6	4	2
14	7	3	5

APPENDIX D. PERMISSION-TO-USE CORRESPONDENCE

From: **Mark Gierl** <mark.gierl@ualberta.ca>
Date: Mon, Mar 8, 2010 at 11:14 AM
Subject: RE: Permission to Adapt a Table from one of your Journal Articles
To: James Masters <jsmaster@uncg.edu>

Good morning Jim,

Yes, you have my permission. Good luck with your dissertation and your upcoming defence.

Mark

From: James Masters [mailto:jsmaster@uncg.edu]
Sent: March 7, 2010 6:42 AM
To: mark.gierl@ualberta.ca
Cc: russelmh@bc.edu
Subject: Permission to Adapt a Table from one of your Journal Articles

Dr. Gierl,

I am finalizing my dissertation under the tutelage of Ric Luecht. The title of my dissertation is A Comparison of Traditional Test Blueprinting and Item Development to Assessment Engineering in a Licensure Context. I have referenced the 2008 JTLA article *Developing a Taxonomy of Item Model Types to Promote Assessment Engineering* co-authored by yourself, Jianwen Zhou, and Cecilia Alves. I would like to adapt table 1 from that article to highlight the model that I am using, which falls under the category of fixed stem, randomly selected options. I sent a request to the editor of the journal about 2 weeks ago, but have not heard back. He may have already forwarded the request to you, but I am defending on March 29th and I need to upload the final version to the graduate school very shortly thereafter. Your assistance in this matter would be greatly appreciated.

Jim Masters
614.798.1457
jsmaster@uncg.edu