A Comparison of Item Parameters across Item Types

Gregory M. Applegate
Purdue University

Follow this and additional works at: https://docs.lib.purdue.edu/open_access_dissertations

Recommended Citation
https://docs.lib.purdue.edu/open_access_dissertations/1497
PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Gregory M. Applegate

Entitled
A COMPARISON OF ITEM PARAMETERS ACROSS ITEM TYPES

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

Deborah Bennett
Ala Samarapungavan

Jerry Gorham
April Ginther

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification/Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy on Integrity in Research" and the use of copyrighted material.

Approved by Major Professor(s): Deborah Bennett

Jerry Gorham

Approved by:

Ala Samarapungavan

10/20/2014

Head of the Department Graduate Program  Date
A COMPARISON OF ITEM PARAMETERS ACROSS ITEM TYPES

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Gregory M. Applegate

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December 2014

Purdue University

West Lafayette, Indiana
This dissertation is dedicated to my family and friends. Their love, support, and encouragement made it possible. In particular, it is dedicated to Denise Bossung whose love, support, and encouragement helped me to focus and complete the process.
ACKNOWLEDGEMENTS

Thank you to my committee members for their advice and guidance.

Thank you to Dr. Deborah Bennett for agreeing to serve as my advisor and committee co-chair, and for shepherding me through the process.

Thank you to my colleagues at Pearson VUE. They have been a wonderful source of education and inspiration.

Thank you to Dr. Karen Sutherland for helping me to understand the content of the examination, and for providing examples of each item type.

A special thank you to Dr. Jerry Gorham for his guidance, knowledge, and wicked sense of humor. I appreciate all three.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
</tr>
<tr>
<td>GLOSSARY</td>
</tr>
<tr>
<td>ABSTRACT</td>
</tr>
</tbody>
</table>

## CHAPTER 1. INTRODUCTION

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Item Formats</td>
<td>4</td>
</tr>
<tr>
<td>1.3 Item Format Effects</td>
<td>11</td>
</tr>
<tr>
<td>1.4 The Measurement Model</td>
<td>18</td>
</tr>
<tr>
<td>1.5 Assessing Items</td>
<td>21</td>
</tr>
<tr>
<td>1.6 Purpose</td>
<td>24</td>
</tr>
<tr>
<td>1.7 Significance</td>
<td>25</td>
</tr>
</tbody>
</table>

## CHAPTER 2. LITERATURE REVIEW

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
</table>

## CHAPTER 3. METHODOLOGY

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Introduction</td>
<td>37</td>
</tr>
<tr>
<td>3.2 Item Sample</td>
<td>39</td>
</tr>
<tr>
<td>3.3 Examinee Sample</td>
<td>42</td>
</tr>
<tr>
<td>3.4 Data Collection</td>
<td>44</td>
</tr>
<tr>
<td>3.5 Sample Preparation and Description</td>
<td>46</td>
</tr>
<tr>
<td>3.6 Analysis</td>
<td>50</td>
</tr>
</tbody>
</table>

## CHAPTER 4. RESULTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Sample Sizes by Item Format</td>
<td>52</td>
</tr>
<tr>
<td>4.2 Responses per Item</td>
<td>53</td>
</tr>
<tr>
<td>Chapter 4.3</td>
<td>Item Difficulty</td>
</tr>
<tr>
<td>-------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>Chapter 4.4</td>
<td>Point Measure Correlation</td>
</tr>
<tr>
<td>Chapter 4.5</td>
<td>Unweighted Standardized Mean Square Fit Value</td>
</tr>
<tr>
<td>Chapter 4.6</td>
<td>Weighted Standardized Mean Square Fit Value</td>
</tr>
<tr>
<td>Chapter 4.7</td>
<td>Response Time</td>
</tr>
</tbody>
</table>

**CHAPTER 5. DISCUSSION**

<table>
<thead>
<tr>
<th>Section 5.1</th>
<th>Introduction</th>
<th>Page 76</th>
</tr>
</thead>
<tbody>
<tr>
<td>Section 5.2</td>
<td>Sample Size</td>
<td>Page 76</td>
</tr>
<tr>
<td>Section 5.3</td>
<td>Item Difficulty</td>
<td>Page 77</td>
</tr>
<tr>
<td>Section 5.4</td>
<td>Point Measure Correlation</td>
<td>Page 78</td>
</tr>
<tr>
<td>Section 5.5</td>
<td>Unweighted and Weighted Standardized Mean Fit Values</td>
<td>Page 79</td>
</tr>
<tr>
<td>Section 5.6</td>
<td>Response Time</td>
<td>Page 80</td>
</tr>
<tr>
<td>Section 5.7</td>
<td>Comparison of Item Formats</td>
<td>Page 83</td>
</tr>
<tr>
<td>Section 5.8</td>
<td>Limitations</td>
<td>Page 86</td>
</tr>
<tr>
<td>Section 5.9</td>
<td>Future Research</td>
<td>Page 88</td>
</tr>
<tr>
<td>Section 5.10</td>
<td>Conclusion</td>
<td>Page 90</td>
</tr>
</tbody>
</table>

**REFERENCES**

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>REFERENCES</td>
<td>93</td>
</tr>
</tbody>
</table>

**VITA**

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>VITA</td>
<td>104</td>
</tr>
</tbody>
</table>
**LIST OF TABLES**

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1: Ethic self-report data for examinees</td>
<td>43</td>
</tr>
<tr>
<td>Table 2: Gender self-report data for examinees</td>
<td>43</td>
</tr>
<tr>
<td>Table 3: Distribution of sample by item format</td>
<td>52</td>
</tr>
<tr>
<td>Table 4: Distribution of sample by item format and content area</td>
<td>53</td>
</tr>
<tr>
<td>Table 5: Descriptive statistics for item difficulty</td>
<td>57</td>
</tr>
<tr>
<td>Table 6: Item difficulty ANOVA results</td>
<td>58</td>
</tr>
<tr>
<td>Table 7: P-values for item difficulty comparisons</td>
<td>58</td>
</tr>
<tr>
<td>Table 8: Descriptive statistics for point measure correlations</td>
<td>60</td>
</tr>
<tr>
<td>Table 9: Point measure correlation ANOVA results</td>
<td>61</td>
</tr>
<tr>
<td>Table 10: P-values for point measure correlation comparisons</td>
<td>62</td>
</tr>
<tr>
<td>Table 11: Descriptive statistics of the unweighted standardized mean square fit value</td>
<td>65</td>
</tr>
<tr>
<td>Table 12: P-values for unweighted standardized mean square fit value comparisons</td>
<td>66</td>
</tr>
<tr>
<td>Table 13: Descriptive statistics of the weighted standardized mean square fit value</td>
<td>69</td>
</tr>
<tr>
<td>Table 14: P-values for weighted standardized mean square fit value comparisons</td>
<td>70</td>
</tr>
<tr>
<td>Table 15: Descriptive statistics for response time</td>
<td>73</td>
</tr>
<tr>
<td>Table 16: P-values for response time comparisons</td>
<td>74</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1: Distribution of response counts by item format</td>
<td>55</td>
</tr>
<tr>
<td>Figure 2: QQ plot for item difficulty</td>
<td>56</td>
</tr>
<tr>
<td>Figure 3: Item difficulty by item format</td>
<td>59</td>
</tr>
<tr>
<td>Figure 4: Point measure correlation by item format</td>
<td>63</td>
</tr>
<tr>
<td>Figure 5: QQ plot for unweighted standardized mean square fit value</td>
<td>64</td>
</tr>
<tr>
<td>Figure 6: Unweighted standardized mean square fit value by item format</td>
<td>67</td>
</tr>
<tr>
<td>Figure 7: QQ plot for weighted standardized mean square fit value</td>
<td>68</td>
</tr>
<tr>
<td>Figure 8: Weighted standardized mean square fit value by item format</td>
<td>71</td>
</tr>
<tr>
<td>Figure 9: QQ plot for response time</td>
<td>72</td>
</tr>
<tr>
<td>Figure 10: Response time by item format</td>
<td>75</td>
</tr>
</tbody>
</table>
GLOSSARY

**Adaptive Test**: Refers to the administration of a test where items are selected based on the ability estimate of the examinee (Thorndike & Thorndike-Christ, 2010, p. 112).

**Construct Irrelevant Variance**: Excess reliable variance that is irrelevant to the interpreted construct (Messick, 1989, p. 34).

**Innovative Item**: An item that makes use of features and functions of the computer to deliver assessments that do things not easily done in traditional paper-and-pencil assessments (Parshall, Harmes, Davey, & Pashley, 2010, p. 215).

**Item**: An instruction or question that requires a response under certain conditions and specific scoring rules (Haladyna, 1997, p. 36).

**Item Difficulty**: The proportion of examinees who get the item correct (Allen & Yen, 1979, p. 121).
**Item Discrimination:** Refers to the degree to which an item differentiates correctly among test-takers in the behavior that the test is designed to measure (Anastasi & Urbini, 1997, p. 179).

**Item Format:** A reference to the form, plan, structure, arrangement, or layout of individual test items, including whether the items require test-takers to select a response from existing alternative responses or to construct a response (Cohen & Swerdlik, 2009, pp. 244–245).

**Item Response Theory:** A family of statistical models used to analyze test data, which provides a unified statistical process for estimating stable characteristics of items and examinees, and defining how these characteristics interact in describing item and test performance (Yen & Fitzpatrick, 2007, p. 111).

**Latent Construct:** A relatively unified underlying trait or characteristic that determines an individual’s ability to succeed with some particular type of cognitive task (Thorndike & Thorndike-Christ, 2010, p. 108).

**Reliability:** Refers to the consistency of measurements when the testing procedure is repeated on a population of individuals or groups (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1999, p. 25).
ABSTRACT

Applegate, Gregory M. Ph.D., Purdue University, December 2014. A Comparison of Item Parameters across Item Types. Major Professors: Debra Bennett and Jerry Gorham.

Seven item formats from a professional licensure examination program (n=4,706) were evaluated to determine whether there are significant differences in item parameter estimates based on item formats. The item formats were standard multiple choice, multiple choice with a graphic or exhibit attached, multiple choice where the examinee was asked to choose the best of four possible correct solutions, multiple choice where examinees were directed to choose the exception, calculations, ordered response and multiple response. Analysis showed similar item parameters for all multiple-choice type items, but better item fit and significantly longer response times for calculation items. Multiple response items were found to produce similar item parameters as multiple-choice items with the exception of increased item difficulty, which is likely due to dichotomous scoring. Ordered response items tended to have longer response times, and are generally more difficult than multiple-choice items. Calculation items were found to be the most discriminating, while ordered response items were the least discriminating.
CHAPTER 1. INTRODUCTION

1.1 Introduction

Determining how much a person knows about a given subject is a complex and difficult process. First, the subject area must be clearly defined. This definition is often referred to as a “latent construct.” The term “latent” refers to the idea that the amount of knowledge a person possesses cannot be directly observed; “construct” refers to the somewhat arbitrary nature of the defined area of knowledge (Allen & Yen, 1979, p. 108; Anastasi & Urbini, 1997, p. 126; Kane, 2006, p. 21). Once the latent construct has been defined, any estimation of ability on the construct depends on observing and interpreting behavior related to the construct. Tests are used to create a situation where behavior related to the latent construct may be observed, and inferences may be made based on those observations. Educational and psychological tests are instruments constructed to measure ability on a latent construct, and are made up of a sample of items that are related to the latent construct.

Items are the first link in the chain of events that connects examinees to the estimation of their ability. Items serve as stimuli to examinees with an opportunity to demonstrate behavior associated with the latent construct. The behavior is then scored according to a prescribed rubric, and the resulting data are used in a measurement
model to estimate the examinee’s ability on the latent construct. Because items serve as the first link in the chain, it is important to understand how their properties may affect subsequent links and the final ability estimate. The purpose of this study is to examine the relationship between item parameters and item format in the context of an item response theory (IRT) measurement model.

Item format refers to the form or layout of a test item, and is a general term that describes the manner in which an item is presented to an examinee. The choice of presentation influences how the examinee interacts with the item, and may influence the cognitive process used to form a response. One example of this is the difference in item difficulty between constructed response and selected response items. The correlation between constructed response and selected response items with identical stems approaches unity (Rodriquez, 2003); however, constructed response items are typically more difficult. This suggests that the items measure the same construct, however, the way in which the item is presented affects the estimation of the item parameter.

Items present an interesting dilemma. They are essential for eliciting observable behavior from examinees; simultaneously, items also affect the quality of examinee responses by introducing construct irrelevant variance into the measurement process. Construct irrelevant variance refers to any reliable variability in the pattern of responses that is unrelated to the latent construct. While construct irrelevant variance is unavoidable, good measurement practice calls for minimizing its effects. Ideally, items should be transparent to the examinee in the sense that the item does not affect the
demonstration of knowledge, skills or abilities. In practice, however, the interaction between examinees and item characteristics affects the validity and reliability of responses. This effect has various causes, including items that use unfamiliar language or symbols, items that depend on unstated cultural assumptions, or item format.

Items consist of a minimum of two parts: 1) an instruction or question that stimulates the examinee to demonstrate observable behavior, and 2) a scoring rubric for assessing how well the examinees’ response demonstrates ability on the latent construct. The focus of this study is on a large-scale professional licensure examination that is presented in a computer-based format. Mass administered written items are typically presented in a selected response format because they allow for efficient and objective scoring. The move from paper and pencil tests to computer-based tests has created an opportunity for the use of several subtypes of selected response items to be used because more complex grading rules may be applied by a computer. In addition to the traditional item format of multiple-choice and fill-in-the-blank, there are new types of items, such as multiple response, ordered response, and multiple true-false formats, to name a few. These new format variations are often called “innovative” or “technology enhanced” items because they may create an opportunity to assess knowledge in new ways. In many cases, the efficacy of these new item formats has not been fully researched.
1.2 Item Formats

The data for this study is derived from a large-scale professional licensure examination. The examination is developed by a national organization that oversees the regulation and licensure of professionals in their field. The item formats studied are all included on the examination.

This study examines seven item formats: the standard multiple-choice item type in addition to three variations of it—two innovative item formats, and one constructed response format. The item formats are multiple choice – standard type (MCS), multiple choice – graphic/exhibit type (MCG), multiple choice – priority type (MCP), multiple choice – follow-up type (MCF), multiple response (MR), ordered response (OR) and fill-in-the-blank calculation (CA).

The multiple choice – standard type (MCS) item format presents an examinee with a stem and four options, and requires the examinee to choose the single correct response. For the purposes of this study, this does not include items with graphics or exhibits, items that ask for a priority decision, or items that ask about follow-up.
MCS Example

A client with newly diagnosed type 1 diabetes mellitus who is being discharged to their home. Which of the following statements by the client would indicate a correct understanding of the teaching?

a. “I should check my glycosylated hemoglobin (HbA1C) level daily.”

b. “I should inject insulin Lispro (Humalog) one hour prior to eating breakfast every day.”

c. “I will recognize thirst, hunger, and frequent urination as signs of hypoglycemia.”

d. “I can use insulin from an opened vial that has been stored at room temperature for up to 30 days.”*

The multiple choice – graphic/exhibit type (MCG) has the same format as a multiple-choice item, but also includes graphics or exhibits to provide additional information to the examinee. While these types of items have been used on paper and pencil tests, the move to computer-based testing has allowed for more detailed and varied materials to be included. MCE items may include line drawings, laboratory results and charts.
MCG Example

**Vital Signs**

Temperature  101.7 ° F  
Pulse  72 beats per minute  
Respirations  26 per minute  
Blood Pressure  124/68 mm Hg

**Miscellaneous Reports**

Electrocardiogram (ECG):

![ECG Image]

**History and Physical**

Pulmonary: crackles in right posterior base, shortness of breath, cough productive of yellow sputum, reports orthopnea for the past 2 nights

Cardiovascular: coronary artery disease, angina, heart failure, hyperlipidemia

Allergies: required intubation after receiving penicillin one year ago

A client who is reporting a cough and fever for the past 3 days. Which of the following prescriptions should be clarified?
a. ceftriaxone (Rocephin) 500 mg, IV, every 12 hours*

b. furosemide (Lasix) 20 mg, IV, once

c. atorvastatin (Lipitor) 20 mg, p.o., daily

d. diltiazem (Cardizem LA) 180 mg, p.o., daily

The multiple choice – priority type (MCP) item format indicates a multiple-choice format where the examinee is asked to make relative comparisons between options. These most often take the form where the examinee is asked to determine which option represents the most urgent need. These item types are intended to measure higher level thinking as described in Bloom’s taxonomy (analysis level).

MCP Example

Four clients have the following laboratory results. Which client should be checked first?

a. with type 1 diabetes mellitus who received regular insulin one hour ago and has a serum blood glucose level of 90 mg/dl

b. with heart failure who receives furosemide daily and has a serum potassium level of 3.3 mEq/liter*

c. who had a total abdominal hysterectomy 8 hours ago and has a white blood cell (WBC) count of 10,000/cu mm.

d. who has end-stage renal disease and has a blood urea nitrogen (BUN) of 70 mg/dl
The multiple choice – follow up type (MCF) item format is a multiple-choice item that is of the exception type where examinees are presented with a situation and four options, and asked to determine which of the four options represents the anomaly or incorrect information. This is similar in concept to a negatively worded item stem, however, no negative words are used in the item. Previous research has shown that negative wording tends to increase item difficulty with construct irrelevant variance (Haladyna & Downing, 1989).

MCF Example

A client has been diagnosed with hypertension. Which of the following statements by the client would require follow-up?

a. “I should avoid adding salt to season my food.”

b. “I will check my blood pressure every morning.”

c. “I should take my prescribed diuretic in the evening before I go to sleep.”*

d. “I will check with my physician prior to beginning an exercise program.”

The multiple response (MR) item format is an item type with a stem and five or six options. There are two to five correct responses. An examinee must choose all of the correct responses, and none of the incorrect responses in order to receive any credit for these items. This type of item has the potential to be scored polytomously, but for the purposes of this study, all items are scored dichotomously since this reflects existing current examination practice. The MR format presents an interesting measurement issue. In theory, multiple response items confound the measurement process by
attempts to measure two or more bits of knowledge simultaneously. An examinee who knows all but one of the correct responses receives the same score as an examinee who chooses all incorrect responses.

MR Example

Which of the following are symptoms of hypothyroidism?

a. Insomnia*

b. hair loss

c. weight loss*

d. tremors*

e. sensitivity to cold

The ordered response (OR) item format presents a set of options which must be placed in the correct order to accurately reflect a common procedure. These are presented as a list of items in one column, which must be moved to a second column in the correct order. Examinees must use each of the options and must place the options in the correct order, in order to receive any credit at all for the item. These items are also scored dichotomously for this study.
OR Example

What is the correct order for changing a sterile dressing for a client with an abdominal wound?

1. Put on sterile gloves.
2. Put on clean gloves.
3. Remove the soiled dressing.
4. Cover the wound with sterile gauze.
5. Remove clean gloves.
6. Clean the wound.

Key = 2, 3, 5, 1, 6, 4

The fill-in-the-blank calculation (CA) item format is a word problem that leads the examinee to a calculation. The examinee must input the correct numerical response in order to receive any credit for the item. These are constructed response type items, and no answer options are provided. Theoretically, CA items should show the best model fit, as they are not as susceptible to random guessing, as are the selected response items.
CA Example

A 6-year-old client who has a prescription for morphine sulfate, 0.1 mg/kg, IV, once. The client weighs 48 pounds. Morphine sulfate 2 mg/ml available. How many ml should be administered?

Key = 1 ml

1.3 Item Format Effects

Isolating effects due directly to format differences is not an easy task. Ideally, two items of identical content and wording would be created in different formats and given to the same group of examinees. The results of the item response analysis could then be compared to determine what differences are related to the change in format since all other factors would be kept constant.

Part of the challenge of this study is that while the items may be drawn from a unidimensional construct, they cannot be matched identically on a specific content task because the formats are, by design, intended to measure different aspects of a concept. For example, an MR item could provide an opportunity for an examinee to demonstrate their understanding of what steps are involved in a specific process, while an OR item would provide an opportunity to demonstrate that the examinee understands the correct ordering of the steps. An attempt to match directly could be made by using five multiple-choice items to determine if the correct order is known (for example: which of the following steps occurs first?), however, this would introduce other differences. It would take five multiple-choice items to compare to one-ordered response item.
Therefore, the set of five multiple-choice items would not be independent, and the probability of guessing the correct response would not be consistent across formats.

But, does item content have to be identical to make valid comparisons? Given a large enough sample, no set of items are truly unidimensional. Multidimensionality could bias the item parameter estimates. Research suggests that estimation of item parameters is robust to violations of unidimensionality in item response theory models (Drasgow & Parsons, 1983; Harrison, 1986; Kirisci, Hsu & Yu, 2001; Reckase, 1979; Wiberg, 2012; Yang 2007).

Reckase (1979) used simulated data to model a dominant latent trait with clusters of items having a secondary, weaker trait. He found that recovery of the original item parameters was robust to minor violations of unidimensionality. Further, Reckase created model data which reflected a highly unidimensional trait with items from an uncorrelated trait. The differences between the two item groups were reflected as poor item discrimination parameters for the uncorrelated trait items.

Drasgow and Parsons (1983) used simulated data to create five data sets that reflected a range of dimensionality from truly unidimensional to unrelated. They found that as long as there was one dominant trait reflected in the items, item parameter estimates were reasonably accurate. Data sets 3 and 8 were specifically designed to reflect a complex construct (for example: knowledge of algebra), and effective recovery of item parameters was possible when the correlation of the dominant trait among items was 0.46 or higher.
Harrison (1986) followed a similar procedure to Drasgow and Parsons, but manipulated the number of items per test and number of common factors. Item parameter recovery worked well with smaller numbers of items, and performed better when there was a single general factor but worked well as long as there was one dominant factor.

Kirisci, Hsu, and Yu (2001) manipulated the distribution of item difficulty (normal, positively skewed or platykurtic) and the number of dimensions (one or three). They found that item distribution did not affect item parameter estimates for tests with more than 40-items, and response samples of more than 1,000. They also suggested using a unidimensional model unless the correlation for the dominant trait is less than 0.4.

Yang (2007) specifically examined the robustness of the Rasch model to violations of the unidimensionality assumption. Yang found that moderate deviations from unidimensionality were modeled by a multidimensional Rasch model (except for item difficulty which showed no differences due to multidimensionality), however, the absolute differences in item parameter estimates were small.

Wiberg (2012) used both real and simulated data, based on a multidimensional college admissions test as a model, to determine if a unidimensional item response model could be used effectively to estimate item parameters for multidimensional data. The real data used five subscales, and the simulated data used two subscales. For the simulated data, the number of items per test was 20 and the number of responses was 500. The correlation between items in different subscales was set to 0.50. She found
that the multidimensional item response theory model exhibited better model fit but that differences in item parameters were small.

The general research consensus on using a unidimensional item response theory model with multidimensional data is that as long as there is a dominant factor in the item content, the presence of multidimensionality has a diminutive effect on item parameter estimations (Wiberg, 2012).

If the measurement model is robust to minor deviations from unidimensionality, this still leaves open the question of: how comparable item parameters are for item format that measures the same content somewhat differently? A good example in the research is a comparison of multiple-choice items and constructed items. These two item formats clearly have at least some level of multidimensionality since multiple choice items require the use of recognition to determine the correct response, while constructed response items require the more cognitively challenging skill of recall.

A meta-analytic research of 67 studies comparing the equivalence of multiple choice (MR) items to constructed response (CR) items found a high correlation between the two formats. When stem equivalent forms are used, a mean correlation of 0.92 was found across studies between MC and CR items. The mean correlation drops to 0.85 when non-equivalent stems are used. Matching items on content made a slight but statistically non-significant increase in the mean correlation for items with non-equivalent stems (Rodriquez, 2003).

Traub (1993) has criticized most studies of the equivalence of MC and CR items as having serious flaws in both design and analysis. He identified nine recent studies
that he felt provided useful analysis, however, he also states that none of the studies fully satisfied his criteria (p. 31). Traub’s conclusion, based on nine selected studies, is that there is evidence to support a difference in the writing domain, contradictory evidence in the reading domain, and no difference in the quantitative domain. He concludes that the best evidence to-date suggests that the two items types do measure the same construct for reading and quantitative knowledge. There is no writing component included in the examination used for this study.

Kuechler and Simkin (2010) have also been critical of the equivalence of MC and CR items. They acknowledge in their literature review that the majority of published research supports the equivalence of MC and CR items but contend that recent studies give reason to doubt the equivalence. They conducted an empirical study with 172 college students in two computer-programming classes. The classes were given tests that consisted of multiple-choice items followed by a constructed response section, which was intended to measure the same information. They divided the MC items, using Bloom’s taxonomy, into three categories: knowledge, comprehension, and application. They analyzed their results by recording the scores for the knowledge MC items, the comprehension MC items, the application MC items, and the CR items. They used linear regression to analyze the results, and reported an adjusted $R^2$ of 0.449. They also calculated Pearson correlation values between each of the four scores. The correlation ranged from a low of 0.078 to a high of 0.875. Additionally, they performed an ANOVA using the MC knowledge levels as independent variables with the CR score as the dependent variable. They found that the mean and standard error for item difficulty
increased monotonically with the level of Bloom’s taxonomy classification. They interpreted this to indicate that student scores varied more widely for higher level MC items.

There are several methodological issues with the Kuechler and Simkin study. They do not report the number of MC or CR items administered as part of the exam, and they state that they gave the MC items first before giving the CR items (it is not clear if students could change their responses after seeing the second part of the test). They do provide an example of what they considered a “complex” MC item but do not report any item parameters to assess the quality of the MC items. They also assumed that items, which are classified as being higher on Bloom’s taxonomy, would be more difficult. However, they present no evidence to support their claim, and other researchers have reported finding no evidence to support this claim (see for example: Seddon, 1978 and Kreitzer & Madaus, 1994 as reported by Haladyna & Rodriguez, 2013 p. 31). Without support for an assumption of linearity, the adjusted $R^2$ statistic is not meaningful.

The correlations that Kuechler and Simkin (2010) calculated are difficult to assess since they do not report the number of items included in the study, and they do not report what percentage of items fell into each level of Bloom’s taxonomy. Moreover, they did not explain what type of rubric was used to score the CR items, nor did they present information to show that the scoring of the CR items was consistent.

The issue of potential multidimensionality due to item format can be viewed from another perspective. Bennett, Rock, and Wang (1991) used data from the
administration of the College Board’s Advance Placement (AP) computer science and chemistry examinations to assess the equivalence of MC item and CR items. They used a two-factor model based on multiple-choice and free-response item formats. They found a single-factor model to be more parsimonious.

Lukhele, Thissen, and Wainer (1994) examined the AP tests for chemistry and U.S. history using a three-parameter item response theory model. They found that multiple-choice items provide about twice the amount of information as free-response items for a given amount of time, and that the free-response items provided little information not already provided by multiple-choice items. Their conclusion was that there is no evidence to indicate that the two different sections measure different constructs.

Given the robustness of the measurement model and the evidence for the validity of the construct measurement across item formats, it is reasonable to conclude that differences in item parameter estimates are likely to be the result of differences in item formats rather than in individual test-taker ability. As a practical matter, if an item format does measure something not included in the definition of the construct, the added measurement noise is considered to be a construct irrelevant variance.

The construct definition for the examination was created by developing a comprehensive job task analysis. Professionals actively working in the field were interviewed and surveyed. The results were reviewed by practitioners and regulators. The analysis was used to develop a consensus definition of the scope of practice for the specific profession. The scope of practice was used by a combination of working
professionals, regulators, and professional test developers to create a construct definition for the examination.

Items for the examination are rigorously developed using best practices, and are reviewed several times by various subject matter experts for content, clarity, and sensitivity. Part of the content review process involves a series of panels, consisting of independent subject matter experts not previously connected to the development process, who review items for a number of traits, including conformity to the construct definition. Psychometric analysis of the items includes estimation of item parameters using a random sample of examinees (typical $n: 400 – 600$). The statistics are anchored to a consistent scale by using current operational test items with known item parameters. Operational items are further analyzed to ensure that item characteristics are stable, and items exhibiting difficulty drift or poor fit are recalibrated before being used again operationally in order to minimize scale drift. All items used for this study followed the process described above to ensure that they are from a unidimensional construct and are measured on a consistent, anchored scale.

1.4 The Measurement Model

The measurement model used for this study is the Rasch model. The Rasch model is the simplest of the family of IRT models. It is a statistical model that uses ordinal level data (counts of correct responses) to develop an interval level measurement by modeling the probability of a correct response (Wright & Stone, 1979). The model estimates two parameters: 1) an ability estimate for each examinee, typically
represented by the Greek letter theta (Θ), and 2) a difficulty estimate for each item, typically represented by the Greek letter delta (δ). Measurements produced by the model are in terms of the natural log-odds of a correct response and are termed logits.

In theory, the range of the measurements produced by the model is negative to positive infinity. In practice, the observed range of most estimates is generally between -3 and +3 logits. The model is relatively simple to use with the count of the correct responses being a sufficient statistic for the estimation process, and the probability of a specific examinee (e) responding correctly to a specific item (i) being modeled by the equation:

\[ P(x_{ei} = 1) = \frac{e^{\theta_e - \delta_i}}{1 + e^{\theta_e - \delta_i}} \]

The accuracy of the model depends on three assumptions. First, there is a monotonic relationship between ability and responses. Second, all items are locally independent of one another in the sense that the response to one item does not influence the response to another item and third, all items are unidimensional. This means that every item used in the model relates to a single latent construct.

The assumption of monotonicity can be checked by calculating the point-measure correlation for each item. Items with a positive point-measure correlation exhibit a monotonic relationship between ability and response level. Any items that have a negative point-measure correlation are excluded from the sample, and the number of items excluded reported by item format.
The second assumption of local independence is met by the item development process. Items are screened during the development process, and items that share common elements are given an enemy relationship in the pool management system. An enemy relationship indicates that the items may have some dependency, and that a response to one item may influence the response to the second item. Items are screened using a computer program that selects items with potentially overlapping content, and the items are then reviewed by content experts. Items with enemy relationships are excluded by the program from appearing on the same test form so each examinee receives a set of items that are locally independent.

The third assumption of unidimensionality is also built into the item development process. The definition for the construct is built by a job task analysis, which defines the scope of practice for entry-level professionals. The content of items is checked to ensure that it complies with the construct definition by a series of reviews, which begins with master level trained professionals who also have training in content development. The process continues with additional reviews by people who are independent of the test development company. These include editorial reviews to ensure that items are presented in a clear and consistent manner, content reviews by working professionals, review of the item content for cultural bias, and reviews by regulators for content and clarity.

The Rasch model was chosen for this study for two reasons: 1) the property of specific objectivity, and 2) because it is the model currently used to develop and maintain the testing program. The examination has approximately 195,000 examinees
per year. The examination is administered as an adaptive test, which means that examinees only see a fraction of all the available items.

The Rasch measurement model possesses the property of specific objectivity (Bond & Fox, 2007; Hambleton & Swaminathan, 1985). In this application, specific objectivity (also referred to as sample invariance) is the concept that the measurement of item properties is independent of the sample used to create the measurement. For the purposes of this study, specific objectivity is a very desirable trait because it allows the comparison of item properties independent of the examinee sample. Eliminating the effects of the sample from parameter estimates eliminates a potentially confounding factor in the data.

Since the Rasch model is used operationally to manage the program, utilizing this measurement model takes advantage of the many data checks built into the process of managing and administering the examination, increasing the fidelity of the study. It also makes it possible to compare items that were administered during various testing cycles since it provides an anchored scale.

1.5 Assessing Items

The extent to which items may contribute to the reliability of ability measurements when using an IRT model is a feature that is typically assessed using item fit statistics. Item fit statistics are measures of how well the data generated by an item meet the expectations of the model. The most common type of fit statistics used to
assess item fit when using the Rasch measurement model is based on a Pearson chi-square approach.

The Pearson chi-square approach uses an analysis of the residual matrix produced by subtracting the expected response predicted by the model, the probability of a correct response along the item characteristic curve, from the actual response. The expected response is a probability with an asymptotic range from 0 to 1, and the actual response is scored as 1 for a correct response or 0 for an incorrect response. Positive residuals result from correct responses, and negative residuals are the result of incorrect responses. The resulting residual matrix will not contain any zero values even for a set of responses that perfectly fits the model.

Because the residual matrix represents the difference between the actual response and the expected response, summing the differences will always result in a value of zero. Therefore, the residuals are squared and the results are used to calculate a mean squares statistic for each item with an expected value of 1 and a range from zero to infinity. This same method may also be used to produce a person fit statistic for each examinee.

There are two versions of the mean square statistic commonly used in Rasch analysis. The first is the unweighted mean square, which is developed as described above. The second is the weighted mean square, which is developed by multiplying each residual by the variance for the item before summing the residuals. The different methods cause the two versions to be interpreted slightly differently. The unweighted mean square tends to be more sensitive to outliers, while the weighted mean squares
places more emphasis on deviations near the difficulty estimate for the item (Bond & Fox, 2007; Wright & Stone, 1979).

The mean square fit statistic is sensitive to sample size (Smith, Schumacker, & Busch, 1998). As sample size increases, it appears that almost all items fit the model. To correct for this, a standardized version of both the weighted and unweighted mean square statistic has been developed. The standardized versions are created using a Wilson-Hilferty transformation of the mean square to obtain t-statistics with an approximately normal distribution (Wright & Stone, 1979). This allows for the interpretation of fit statistics using the familiar normal curve theory where a perfectly fitting item would have a value of 0, and values above or below 0 would indicate the number of standard deviations away from the mean. The standardized statistics also have a flaw in that the rejection rate increases as the sample size increases (Bond & Fox, 2007).

One of the advantages of using the standardized weighted and unweighted mean squares is that they provide a measure of both underfit and overfit. Typically, a critical value of plus or minus 2 to 3 is used to evaluate fit (Bond & Fox, 2007). Underfit is represented by positive values that exceed the specified critical value. These are items that have more variation in the response matrix than are predicted by the model. Underfit may be caused by guessing, unclear items, double-keyed items or mis-keyed items. Overfit is represented by negative values below the specified critical value. These are items that have less variation in the response matrix than would be predicted by the model. Overfit may be caused by cueing, multidimensionality, or differential item
functioning. Both underfit and overfit are important considerations as both types of fit measures are indicative of problems which suggest that the Rasch model may be inappropriate for the type of data (Masters, 1988).

One concern when estimating fit statistics is how well they can be estimated when based on a sparse matrix that may be routinely produced by an adaptive test. On an adaptive test, candidates only see a select few of the total number of items available. For operational items, this is controlled by an algorithm that attempts to match an item difficulty with the most recent ability estimate for the examinee. For experimental items, such as the ones used in this study, the items are chosen at random from the available pool so that all levels of examinees see the item. In both cases, there exists a large data matrix where the majority of examinees do not see the majority of items. Research using Rasch fit measures, both the standardized and unstandardized weighted and unweighted mean squares and the point-measure correlation, has found that estimates based on the sparse data matrix have relatively high fidelity (Wolfe & McGill, 2011).

1.6 Purpose

The purpose of this study is to explore differences in item parameters associated with different item formats in examinations. Developing an understanding of how item formats affect item parameters provides information that can reduce the standard error of measurement for test scores and reduce bias in testing. Items are imperfect measurement instruments, and all items have some level of construct irrelevant variance to test scores. Item format can also affect the response time needed by the
test-taker and, along with the scoring rubric, this affects how much information is
gathered about the test-taker’s ability.

The goal of this study is not to label particular item formats as “good” or “bad”
but to begin to develop an understanding of the trade-offs involved in using one item
format rather than another. Ideally, this study will find that there is no significant
difference in item parameters due to item format. This would provide item developers
with the most flexibility as they create new tests.

1.7 Significance

Often times, item content can be presented in several different item formats.
Test developers must make decisions about which item format to use. Some formats
may provide relatively better item discrimination or model fit, which reduces the
standard error of measurement, but may require a longer response time, which limits
the amount of content which can be tested. This is particularly true in the specific case
of adaptive tests where differences in item difficulty can change the types of items a
candidate receives during the examination.

Previous research has shown that not all item formats produce similar item
parameters. Multiple-choice items and constructed response items produce very similar
ordering of candidates, however, they produce significantly different item difficulties
(Rodriguez, 2003). The type K multiple-choice item format was designed to better
measure critical thinking with a selected response item format, however, research
found that it created significantly more construct irrelevant variance compared to other
item formats (Albanese, 1993). Alternative choice items can be used to test a broad sample of content within a short period of time, however, poor discrimination makes using this item format of questionable value (Downing, Baranowski, Grosso, & Norcini, 1995).

Seven item formats and five item parameters are included in this study. The null hypothesis for the study is that all item formats studied will produce item parameters with similar values.
CHAPTER 2. LITERATURE REVIEW

The importance of creating high quality items and, in particular, high quality selected response items for testing has been of interest to researchers throughout the 20th century and into the 21st century, though the research and level of interest has been somewhat inconsistent. Several researchers have noted that item development is an understudied topic in psychometrics (Cronbach, 1970; Haladyna, Downing & Rodriguez, 2002; Nitko, 1985; Roid & Haladyna, 1982). A thorough examination of the various editions of Educational Measurement supports this point of view. The first two editions of Educational Measurement each devoted a chapter to item development (Ebel, 1951; Wesman, 1971), while the most recent two editions only touch upon the subject in relation to other psychometric concepts (Brennan, 2006; Linn, 1989).

Despite the lack of research emphasis in the psychometric community, item writing is an important topic, and one of the foundations for developing useful tests. The negative effect of poorly written items on the measurement of examinee ability has been documented (Downing, 2002, 2005; Tarrant & Ware, 2008), and other studies have demonstrated an inverse relationship between item discrimination and specific item characteristics. Discrimination refers to the ability to separate examinees with low ability from examinees with high ability (Thorndike & Thorndike-Christ, 2010).
There are several examples of item formats that have a negative effect on item parameters. The complex multiple-choice format, where an examinee is presented with a stem and two sets of options—the second set of options being combinations of the first set, and asked to select an option from the second set—has been shown to be less discriminating than the standard multiple-choice format (Albanese, 1993). Another example is the use of an inclusive option in selected response items. Use of an inclusive option reduced item discrimination when compared to the same stem with specific options (Dudycha & Carpenter, 1973). As a final example, negatively phrased items have been shown to reduce item discrimination (Harasym, Doran, Brant & Lorscheider, 1993) and increase item difficulty (Cassels & Johnstone, 1984).

One reason that researchers may be reluctant to address the issue of item development is the opinion held by some members of the psychometric community that item writing is more art than science (Downing, 2005; Rodriquez, 1997; Wood, 1977). However, other researchers have been exploring the idea of computer-generated items that require a minimum of human processing (Bejar, 2010; Embretson & Yang, 2007; Wainer, 2002), suggesting that there are at least some recognizable and applicable rules for developing high quality items.

This topic is currently being investigated by several researchers. It seems likely that there are at least some characteristics of good items that can be quantified and applied generally (see for example: DiBattista, Sinnige-Egger, & Fortuna, 2014; Haladyna & Rodriguez, 2013; Gierl & Haladyna, 2012; or Stopek, Eve, & Burke, 2014).
Having clear guidelines for developing effective test items is crucial because poorly constructed items introduce construct irrelevant variance as has been demonstrated by a negative effect on item discrimination (Richichi, 1996) and the estimation of ability (Downing, 2005, Tarrant & Ware, 2008). Construct irrelevant variance is error introduced into the measurement from factors unrelated to the construct being measured (Haladyna & Downing, 2004). A good example of construct irrelevant variance is the inclusion of cultural references in items that are designed to measure a construct which does not have a cultural understanding component. A mathematics question using pizza as an example of a circle may function well for most examinees, but could cause examinees that are not familiar with pizza to have difficulty understanding the intent of the item. The concept of pizza is unrelated to the construct of mathematical ability, and unfamiliarity with pizza introduces systematic error into the measure of mathematical ability.

Current best practices for item development consist of a set of 31 guidelines based upon existing research and textbook recommendations (Haladyna, 2004; Haladyna, Downing & Rodriquez, 2002). The guidelines were generally compiled from research-based consensus methods, but are not equal in importance. Some of the guidelines are supported by research while others are simply the result of a survey of psychometric texts. Item development is, thus, an immature technology and more research needs to be done. Presently, the set of 31 item-writing guidelines represent the state-of-the-art in item development.
Developing effective items is highly dependent upon language—both the language used in the item, and the examinee’s understanding of the language and its cultural assumptions. However, the issue of language use is not clear-cut. While poorly phrased or unfamiliar language may obscure the intent of an item, knowledge of specific terms is a part of many constructs. Some changes in wording may result in clearer measurement while others may adversely affect item validity.

For example, studies involving examinees in 10th grade or younger have found significant differences in item difficulty estimates when phrasing changes are made on items (Benson & Crocker, 1979; Bolden & Stoddard, 1980) while studies involving examinees in the 11th grade and older do not find corresponding significant changes (Bornstein & Chamberlain, 1970; Green, 1984). The decrease in the impact of phrasing changes on item difficulty as the age of the test sample increases is consistent with a single study that used a sample of 4th grade through college age students to study variables related to item difficulty (Jerman & Mirman, 1974). A possible explanation for this difference is that once a given threshold of reading comprehension has been achieved, changes in phrasing no longer produce significant effects (Green, 1984).

Similarly, Cassels and Johnstone (1984) found that the substitution of key words with simpler language, and changing an item stem from negative to positive wording resulted in changes in item difficulty. They also found that when items comprised of complex sentences were modified to simple sentences, item parameters were affected. However, minor changes in the parts of speech or framing the item in terms of passive or active voice did not appear to affect item parameters. The study sample was
comprised of approximately 3,600 students around the age of 16. Their conclusion was that changes which do not affect short-term memory requirements, the clarity of the item, or the thinking process needed to respond correctly to an item do not affect item parameters. In some cases, they substituted scientific terms with their definitions resulting in a reduction in item difficulty. The decrease in difficulty appears to be more related to a lower level of the construct rather than an issue of understanding language in general. Because of this issue, it can be difficult to interpret readability indices, particularly for professional licensure or certification tests which include specialized vocabulary.

More directly to the point of this study, two item formats have been shown to inherently introduce construct irrelevant variance beyond what is normally expected. The complex multiple-choice format, most commonly seen as a Type K format, was popular in medical licensure and certification examinations during the 1970s and 1980s (Rodriquez, 1997). Developed by the National Board of Medical Examiners as a method to test higher order thinking (Hubbard, 1978), the format consists of a stem, a set of options, and then a second set of options that represented selected combinations of the first set of options. The format was found to introduce an element of cueing that allowed some examinees to respond correctly to an item without knowing the content, reducing the reliability of test scores (Albanese, 1993; Albanese, Kent & Whitney, 1979; Harasym, Norris & Lorscheider, 1980; Kolstad, Briggs, Bryant & Kolstad, 1983).
Multiple Choice Type K Example

A client with newly diagnosed type 1 diabetes mellitus is being discharged to their home. Which of the following statements by the client would indicate a correct understanding of the teaching?

a. “I should check my glycosylated hemoglobin (HbA1C) level daily.”
b. “I should inject insulin Lispro (Humalog) one hour prior to eating breakfast every day.”
c. “I will recognize thirst, hunger, and frequent urination as signs of hypoglycemia.”
d. “I can use insulin from an opened vial that has been stored at room temperature for up to 30 days.”

i. Both A and B ii. D only iii. Both B and C iv. C only

Another more subtle example of item format affecting measurement is the use of negative wording. Negative phrasing has been used in surveys and inventories as a check of reliability, however items using negative phrasing (e.g. not or except) have been shown to measure something different than positively worded items (Dudycha & Carpenter, 1973; Schriesheim & Eisenbach, 1995; Spector, Van Katwyk, Brannick & Chen, 1997; Stewart & Frye, 2004). This effect may be related to item readability as other research has found that it disappears in items of lower reading difficulty (Tamir, 1993), or with test-takers with high reading ability (Downing, Baranowski, Grosso & Norcini, 1995). The additional difficulty of the task may be adding construct irrelevant variance to the item.

Previous research on item characteristics has focused on changes in item difficulty and discrimination (see for example Albanese, 1993; Ascalon, Meyers, Davis, &
Smits, 2007; Casler, 1983; Crehan & Haladyna, 1991; Frary, 1991; Green, 1984). Using changes in item difficulty as a measure of item quality is problematic. A change in item difficulty that is entirely related to the latent construct would not change the measure of the examinee’s ability; it would merely change the probability that a specific examinee would respond correctly.

Changes in item difficulty are only related to item quality when they reflect the measurement of abilities outside the latent construct. For example, it has been demonstrated that item difficulty changes when definitions replace terms in science items (Cassels & Johnstone, 1984). The reduction in item difficulty does not signify a greater understanding of the construct but merely a greater understanding of the specific item. Comprehension of terms is an important part of any discipline and if items containing terms are more difficult, it is likely because they require a greater ability in the construct. An increase in item difficulty by itself does not indicate a lower quality item. However, if the change in item difficulty is due to a cueing effect, it could affect the ordering of examinees and degrade measurement.

While item difficulty is not a good indicator of item quality, it is useful for comparing item formats. If two formats have similar model fit and discrimination but differ significantly in item difficulty, this knowledge could be used to help build more robust item banks by allowing for the creation of items with a desired distribution of difficulty values.

Discrimination measures in item research have been more effective at quantifying item quality. Various statistics have been used in classical test theory to
estimate discrimination including D, phi, biserial ρ, and tetrachoric ρ (Crocker & Algina, 1986). One of the most commonly used measures has been the point-biserial correlation which may be applied separately to the item key and to each item distractor.

Items with greater discrimination reflect a better separation between ability levels, however relying upon discrimination as a measure of item quality has its limits. Increased discrimination may be caused by a number of factors not all of which are a function of ability within the latent construct (Masters, 1988). Issues such as cueing, or interdependence, where one item provides information helpful to responding to another item, may increase a discrimination estimate leading to the erroneous conclusion that an item is effectively measuring the construct.

Item fit statistics developed from item response models are a more effective tool for evaluating how well an item functions within a measurement model. Item fit statistics provide a measure of how suitably an item conforms to the expectations of the measurement model in terms of both under-fitting and over-fitting the model assumptions (Bond & Fox, 2007, Wright & Stone, 1979). Items that underfit the model have more variation in the response matrix than is predicted. Underfit may be caused by guessing, unclear items, double-keyed items or mis-keyed items. Items that overfit the model have less variation in the response matrix than is predicted. Overfit may be caused by cueing, multidimensionality, differential item functioning, or other factors.

Researchers have found that using the standardized mean squares statistic is more reliable than using the mean squares if an appropriate critical value is chosen (Smith & Suh, 2003). However, standardized Rasch fit statistics have been criticized by a
number of researchers (e.g. Busmeyer, 1980; Krantz & Tversky, 1971; Nickerson & McClelland, 1984; Nygren, 1980). Much of this criticism is based on the null distribution for Rasch fit statistics not being constant across sample sizes (e.g. Molenaar & Hoijtink, 1990: Smith, 1991). Variations to the null distribution are associated with the number of items analyzed, and the sample size of examinees.

Molenaar & Hoijtink’s (1990) research examined fit from a person fit perspective but, for the Rasch model person and item fit, and found that statistics are mathematically symmetrical and conclusions about distributional properties are similar. They examined the null distribution by simulating data and comparing the skewness and kurtosis to a unit normal curve. They found that the distribution of fit statistics are biased (negatively skewed and leptokurtic) but that the effects decrease as sample size increases, and also when the items are similar in difficulty to the mean ability level of the sample. Furthermore, the bias effect is small for sample sizes over 30. Calculating the standard error of skewness and the standard error of kurtosis for each sample level used in the study shows that few of the conditions would be flagged as a non-normal distribution using a critical value of 2.

Wang & Chen (2005) developed a correction for the standardized fit statistics. It adjusts the critical value by the difficulty estimate to obtain a more stable and predictable distribution. Their research showed that the correction was useful but only has a practical effect when the number of items analyzed is less than 20.

Smith (1991) approached the issue from a slightly different perspective by calculating the mean and standard deviation across variations in sample size and
number of items. He also found differences in the null distribution, but again the differences were closely related to sample size and number of items. Smith found that the mean and standard deviation increased with an increase in sample size, however the effects were reduced as the number of items analyzed increased. The effect is that as sample size increases, standardized fit statistics become more sensitive, leading to a higher rejection rate for items.

While it is clear that the null distribution for standardized fit statistics is not constant, practical problems with using the standardized fit statistics only occur when the number of items being analyzed is small, the sample size is small, or the sample size is very large. A small number of items being analyzed, or a small sample size result in an item rejection rate that is smaller than expected, while large sample sizes result in an item rejection rate that is higher than expected.

Ideally, this study will show that item format has no effect on the fit of items. The lack of a difference would indicate that all the formats studied may be useful for constructing tests, and the choice of format could be chosen in relation to the type of content being tested. If it is found that item format does indeed affect fit, this would suggest that some item formats are more susceptible to introducing construct irrelevant variance and need to be used with care. The particular item formats chosen for this study were selected because they are common to many large-scale testing programs, and information regarding the fit of these item formats would have broad applicability.
CHAPTER 3. METHODOLOGY

3.1 Introduction

The goal of this study is to determine whether item format alone significantly changes item parameter estimates. To determine this, response data were collected on seven item formats, and evaluated using five fit measures. The data were analyzed using analysis of variance to detect differences in mean fit estimates. The null hypothesis for this study is that there will be no significant differences on average standardized mean squares across item type formats.

The items and data used for this study are from an internationally administered professional licensure examination. The examination is administered as a computerized adaptive test (CAT). Administering the examination as a CAT provides better item security, and allows for shorter examinations by only administering items that are appropriate to a given examinee’s ability. Each time an examinee responds to an item, a new estimate of the candidate’s ability is generated. Based on the ability estimate, the next item is selected so that the examinee has approximately a 50% chance of responding correctly to the item. Using this method maximizes the amount of information provided by each item. This process is repeated until a pass/fail decision can be made. A pass/fail decision is made when: 1) an examinee has completed the
minimum number of items and their ability estimate is two standard errors above or below the cut score, 2) an examinee has run out of time (6 hours), or 3) an examinee has completed 265 items.

On average, an examinee will respond to 119 items, and complete the examination in 2 hours and 18 minutes. Slightly over half (52%) of examinees will complete the examination by responding to only the minimum number of items (75), and roughly 13% of the examinees will respond to the maximum number of items (265). Approximately 1% of candidates use the maximum amount of time allowed (6 hours). The passing standard at the time of data collection was -0.16 logits, and the examination has an estimated decision consistency of 0.92. The mean theta estimate for U.S. educated, first time test-takers is 0.395.

The estimated decision consistency for the examination is a measure of reliability. It is a statistical estimate, using a single administration of the examination, which represents the likelihood that an examinee would be classified the same if given a second administration of the examination (Stearns & Smith, 2008). Decision consistency is improved when items are well targeted to the examinee and/or more items are presented, however improvements are asymptotic as the estimate approaches 1.

The high-stakes nature and significant cost of this examination creates a situation where examinees are motivated and focused on completing the examination to the best of their ability. The examination is administered as a computer-based test continuously during the year in professional test centers. These test centers employ a number of security protocols to ensure the integrity of the testing process. These
protocols include using biometric information for examinee identification, digital audio and video surveillance of examinees during the testing period, and live proctors to monitor examinees as they work in partitioned workstations.

Examinees must qualify to take the test by completing a collegiate level course of study (typically a Bachelor of Science degree but at a minimum an Associate Degree). The course of study includes academic coursework and clinical experience. Examinees must first receive authorization to test from the appropriate regulatory agency (typically at the state level), and must present evidence that all prerequisites for testing have been satisfied. Accommodations are available for examinees with disabilities upon approval of the regulatory agency. The most common accommodation requested and granted is extra time.

Test results are analyzed using the Rasch model, and results are reported to examinees on a pass/fail basis. A single cut score is used to make the pass/fail decision. Items are administered adaptively with a minimum of 75 and a maximum of 265 items being presented within a six-hour time limit. Examinees are free to take breaks during the examination with the break time counting towards the six-hour time limit. Test proctors are available to help with questions about administration, or to provide technical support but the proctors are specifically prohibited from discussing items.

3.2 Item Sample

The items for this program are developed using current best practices in the field of test development. The item development program includes a style guide to maintain
consistent item presentation, and to maximize item clarity. Items are created by professional educators from across the United States during item writing sessions held in the test development offices in Chicago, Illinois. The items developed during the sessions are then given to item developers who verify and document both the correctness of the key and the incorrectness of each distractor by referencing educational textbooks. The items are then edited for style and clarity.

After the items have been referenced and edited, they are sent to editorial staffs who verify that item changes are in accordance with the program style guide. This includes ensuring that each item has proper spelling and grammar, uses professional terms and acronyms consistently, and that all measurements and numbers are rendered consistently. It also includes removing extraneous or irrelevant information from an item including unnecessary references to gender, ethnicity or age. After the editorial work is complete, the items are reviewed by an outside group of subject matter experts who confirm that the items are clear, correct and appropriate. The items are then sent to a final review. Items completing this process are then eligible to be included on a live examination as experimental items.

Experimental items included on a live examination are not scored. They are intermingled with operational items but are not separately identified. Examinees are informed that experimental items are included on an examination but the items are not separately identified. The number of experimental items presented to each examinee is consistent, but otherwise experimental items are administered randomly. Experimental items are administered without regard to examinee ability estimates.
Experimental items are administered based on a uniform probability distribution without replacement. Given that there are x experimental items available, the probability of an examinee receiving a specific experimental item is 1/x. The probability of the next experimental item being chosen is 1/(x-1) because the first item chosen is no longer available to be presented. This process continues until the maximum number of experimental items has been reached. The administration algorithm is set so that no examinee receives more than two consecutive experimental items, and all the experimental items are administered before the minimum number of items has been reached. This ensures that each complete examination contains the same number of experimental items.

Because items are administered randomly based on a uniform distribution, the number of examinees who respond to each experimental item will be normally distributed (central limit theorem). Given a large sample size from a population with a finite variance, the mean of the sample will be approximately equal to the mean of the population, and the sampling distribution of the mean will be normally distributed.

The order of item presentation varies for each examinee. Items are presented one at a time; a response must be given before the next item is presented, and no review of previous items is allowed. A timer is displayed on the computer screen showing examinees how much time is left for the examination. Because experimental items are presented early in the examination, time constraints should have a minimal effect on examinees’ response pattern.
The sampling size for each experimental item is dependent on the number of examinees who test during each deployment of experimental items. Examinee volumes are relatively consistent but cannot be predicted precisely. The target sample size for items in this program is 525 examinees per item, and the range is typically between 400 and 600 examinees. The number of items included in the study is 4,706.

3.3 Examinee Sample

Each year, approximately 195,000 examinees take the examination. Because it is an adaptive test, not all examinees receive the same items. It is not practical to create a specific profile of each sample for each item so a general profile of examinee characteristics is presented. The profile presented is based on demographic information from candidates who took the examination during the period that the data for this study were collected. Only results from examinees educated in the United States, who are taking the examination for the first time, were used for the calibrations.
Table 1: Ethnic Self-report Data for Examinees

<table>
<thead>
<tr>
<th>Ethnic Group</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>African American</td>
<td>9.1</td>
</tr>
<tr>
<td>Asian</td>
<td>10.7</td>
</tr>
<tr>
<td>Caucasian</td>
<td>55.6</td>
</tr>
<tr>
<td>Hispanic</td>
<td>6.1</td>
</tr>
<tr>
<td>Native American</td>
<td>0.5</td>
</tr>
<tr>
<td>Pacific Islander</td>
<td>1.2</td>
</tr>
<tr>
<td>Other</td>
<td>5.0</td>
</tr>
<tr>
<td>Not reported</td>
<td>11.8</td>
</tr>
</tbody>
</table>

Table 2: Gender Self-report Data for Examinees

<table>
<thead>
<tr>
<th>Gender</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>83.6</td>
</tr>
<tr>
<td>Male</td>
<td>13.7</td>
</tr>
<tr>
<td>Not reported</td>
<td>2.7</td>
</tr>
</tbody>
</table>

Approximately 87% of the sample group passes the examination. Examinees respond to an average of 112 items each. Mean testing time is approximately two hours. Less than one percent of the sample examinees use the maximum amount of time allowed. About one-half of the examinees take a break during the examination period.
3.4 Data Collection

Data were collected from several test deployments. The response matrix was analyzed using a computer program designed to perform Rasch analysis (Winsteps). Item difficulty was anchored to the current operational scale used by the testing program. Anchoring items by difficulty does not affect measures of item fit. The dependent variables were calculated using the Winsteps program based on the response matrix, and the independent variables are based on coding by subject matter experts.

**Dependent Variables**

1. Unweighted standardized mean square fit value—This is the traditional Pearson chi-square fit statistic commonly used to evaluate model fit.
2. Weighted standardized mean square fit value—This is the traditional Pearson chi-square fit statistic weighted by the information function for each item.
3. Point-measure correlation—The Pearson product moment correlation between the test score and the examinee ability estimate. It is similar to the point-biserial correlation, and is interpreted similarly.
4. Item Difficulty—The delta estimate of the difficulty of the item on a logit scale. Differences do not indicate greater or lesser item quality, but may be useful for creating items that match a specific difficulty profile.
5. Response Time—The amount of time an examinee had the item displayed on the computer screen before advancing to the next item.
Independent Variable

The independent variable is item format. The levels of the independent variable are as follows:

1. Multiple-choice item format / standard type (MCS)—a binary variable which indicates a multiple-choice item that is of the standard type where examinees are presented with a stem and four options, and are asked to choose the single correct response. This category does not include items with graphics/exhibits, or items that ask for a priority decision, or ask about follow up.

2. Multiple-choice item format / graphic/exhibit type (MCG)—a binary variable indicating that the item includes graphics or exhibits to provide additional information to the examinee.

3. Multiple-choice item format / priority type (MCP)—a binary variable which indicates a multiple-choice format item that is of the priority type where examinees are presented with four options and asked to determine which option represents the most urgent need. These typically measure an examinee’s ability to make distinction between options and judgments following a general set of guidelines.

4. Multiple-choice item format / follow up type (MCF)—a binary variable which indicates a multiple-choice item that is of the exception type where examinees are presented with a situation and four options, and are asked to determine which of the four options represents the anomaly or incorrect
information. This is similar in concept to a negatively worded item stem, however, no negative words are used in the item.

5. Multiple response item format (MR)—a binary variable which indicates an item type with a stem and five or six options. There are two to five correct responses, and the items are scored dichotomously. An examinee must choose all of the correct responses and none of the incorrect responses in order to receive credit for these items.

6. Ordered response item format (OR)—a binary variable which indicates a set of options which must be placed in the correct order to model a procedure. These are presented as a list of items in one column which must be moved to a second column in the correct order. Examinees must use all of the options and have them in the correct order to receive credit for these items.

7. Fill-in-the-blank calculation format (CA)—a binary variable which indicates a word problem that leads the examinee to a calculation. The examinee must input the correct numerical response to receive credit for these items. These are constructed response type items, and no answer options are provided.

3.5 Sample Preparation and Description

The following information was collected for each item in the study:

1. Item Format (MCS, MCE, MCP, MCF, MR, OR or CA)
2. Number of examinees who responded to the item
3. Delta parameter (item difficulty on a logit scale)
4. Standardized unweighted mean square fit statistic

5. Standardized weighted mean square fit statistic

6. Point-measure correlation statistic

7. Response Time

The initial data screening eliminated any items that met any of the following conditions.

1. Items with missing values

2. Items with responses from less than 400 examinees

The sample size, mean, median, standard deviation, minimum and maximum statistics were calculated for each dependent variable by item format. The values were checked to ensure they were within reasonable ranges. Descriptive statistics by format are provided.

Operational items were administered adaptively, however the sample items for this study were administered using a simple random selection method without regard to the test-takers’ ability estimate. Each test-taker received 15 of the sample items. The sample items were selected using a uniform probability distribution without replacement.

The mean theta estimate for the calibration sample is 0.395. Warm up and fatigue effects were minimized by allowing for ten operational items to be administered
before any sample items were administered, and by ensuring that all sample items were administered within the first 75 items.

The point-measure correlation is used to measure discrimination for this study rather than the more common point-biserial correlation. While the point-biserial correlation works well for items of moderate difficulty, it is less effective for items at the extremes of item difficulty. This is due to a ceiling effect that limits the value for item difficulty from 0 to 1. An alternative when using item response test theory is the point-measure correlation. The point-measure correlation is a Pearson product moment correlation using test scores and examinee ability estimates. Because examinee ability estimates have a wider range and are on an interval scale, the point-measure correlation does not suffer as much from the ceiling effect. In practice, the point-biserial correlation and the point-measure correlation are usually similar, and are interpreted in the same way. The point-measure correlation is used for the purposes of this study because its estimation has been shown to be consistent across difficulty levels when developed using a sparse matrix similar to the type of data used in this study (Wolfe & McGill, 2011). Both statistics have limitations when used for traditional distractor analysis (Attali & Fraenkel, 2000).

Given the nature of this exam, the point-measure correlations may be challenging to interpret. A variety of factors can cause discrimination estimates, particularly for a high-stakes examination with groups of test-takers who have similar backgrounds and abilities. Better estimates of discrimination could be obtained by using a broader sample. By including test-takers with a wide range of training, better
estimates of true item discrimination could be developed. The sample of test-takers used to calibrate the items for this study have all passed a college-level curriculum to prepare them for the examination, and in many cases have taken a practice exam to ensure that they are ready. This imposes a range restriction on the point-measure correlation that limits the estimates of discrimination (Haladyna, 2004, p. 211).

Another reason the point measure correlation estimates may be relatively low and clustered close together is the nature of the examinee population. Examinees come from thousands of programs across the United States. While there are curricular guidelines for the professional programs, there is no required standard curriculum. In addition to differences in curriculum, there are literally hundreds of different textbooks used in the professional programs throughout the United States. These textbooks sometimes have variant or conflicting information for particular processes and procedures.

A third factor that may contribute to low point-measure correlation estimates is the rapid pace of change within the profession. Practice changes are based on the latest medical research. Each year, many items are removed from the item pool because they do not reflect current practice. Changes in practice sometimes take years to be fully adopted by the profession, while the examination always reflects the most current changes to practice. As a result, differences in candidate ability may be confounded with differences in curriculum, teaching and changes in practice.

For purposes of this study, the important issue is not the absolute value of the point-measure correlations, but the relative values. The primary goal is to find
differences in item discrimination, and the relative measures are sufficient for this purpose.

3.6 Analysis

The analysis of variance (ANOVA) method was used to determine whether a significant difference exists between the means of the item formats for each of the item parameters. A separate ANOVA procedure was run for each dependent variable. If the ANOVA produced an insignificant F test ($\alpha = 0.01$), the analysis was complete (note that the overall alpha for analysis is 0.02, but a Bonferonni adjustment was made to account for the multiple ANOVAS). Assumptions underlying the ANOVA methods were examined for each variable to determine if the ANOVA was appropriate. Since the sample sizes for each item format are significantly different, an unbalanced ANOVA design, using pool variances, was used for the analysis.

In cases where the F test was significant, a post-hoc analysis was conducted. A Tukey-Kramer test was used for the pairwise comparisons. The Tukey-Kramer method controls the familywise error rate for unequal sample sizes. All possible pairs were compared. The R-square statistic was not interpreted for this study since there is no assumption of linearity based on the nominal categories of independent variables.

The Type I error rate for an ANOVA is represented by the alpha level (0.01 for this study). The Type II error rate is represented by the beta level. The beta level or power of the statistical test is based on the sample size used for the analysis. The data
for this sample have unequal sample sizes so the smallest sample size was used for the power calculation.

When the assumptions for the ANOVA were not met, a non-parametric test—the Kruskal-Wallis test—was used to compare means. When the F test was significant, a post-hoc analysis of all pairwise comparisons was conducted. To control for familywise error, a Bonferroni adjustment was made. Given the alpha level of 0.01 for this study and the 21 possible pairs, the adjusted Bonferroni significance level is 0.000475.
CHAPTER 4. RESULTS

4.1 Sample Sizes by Item Format

The total number of items analyzed was 4,706. Table 3 below shows the distribution of items by format.

Table 3: Distribution of Sample by Item Format

<table>
<thead>
<tr>
<th>Item Format</th>
<th>Frequency</th>
<th>Percent</th>
<th>Cumulative Frequency</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calculation</td>
<td>64</td>
<td>1.36</td>
<td>64</td>
<td>1.36</td>
</tr>
<tr>
<td>Multiple Choice Follow-up</td>
<td>76</td>
<td>1.61</td>
<td>140</td>
<td>2.97</td>
</tr>
<tr>
<td>Multiple Choice Graphic/Exhibit</td>
<td>190</td>
<td>4.04</td>
<td>330</td>
<td>7.01</td>
</tr>
<tr>
<td>Multiple Choice Priority</td>
<td>70</td>
<td>1.49</td>
<td>400</td>
<td>8.50</td>
</tr>
<tr>
<td>Multiple Choice Standard</td>
<td>3,330</td>
<td>70.76</td>
<td>3,730</td>
<td>79.26</td>
</tr>
<tr>
<td>Multiple Response</td>
<td>845</td>
<td>17.96</td>
<td>4,575</td>
<td>97.22</td>
</tr>
<tr>
<td>Ordered Response</td>
<td>131</td>
<td>2.78</td>
<td>4,706</td>
<td>100.00</td>
</tr>
</tbody>
</table>

The smallest sample size for this study is 64. The power analysis shows that the beta level for an effect size of 0.50 standard deviations is 0.26, and for an effect size of 1.00
standard deviations is 0.968. Effect sizes of 1.25 standard deviations have a beta level of 0.999 or better.

Table 4 below shows the distribution of item formats within each of the eight subcategories on the test. With the exception of calculation items, which are only represented in areas 5 and 6, the percentage of each item format in each category is roughly similar.

<table>
<thead>
<tr>
<th>Item Format</th>
<th>Area 1</th>
<th>Area 2</th>
<th>Area 3</th>
<th>Area 4</th>
<th>Area 5</th>
<th>Area 6</th>
<th>Area 7</th>
<th>Area 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calculation</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>6.3%</td>
<td>2.3%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Multiple Choice Follow-up</td>
<td>1.0%</td>
<td>1.1%</td>
<td>2.6%</td>
<td>1.1%</td>
<td>1.4%</td>
<td>1.2%</td>
<td>2.6%</td>
<td>1.8%</td>
</tr>
<tr>
<td>Multiple Choice Graphic/Exhibit</td>
<td>0.5%</td>
<td>2.9%</td>
<td>2.6%</td>
<td>0.2%</td>
<td>2.9%</td>
<td>3.9%</td>
<td>3.0%</td>
<td>10.4%</td>
</tr>
<tr>
<td>Multiple Choice Priority</td>
<td>2.9%</td>
<td>0.0%</td>
<td>1.0%</td>
<td>0.5%</td>
<td>0.2%</td>
<td>0.9%</td>
<td>2.6%</td>
<td>2.9%</td>
</tr>
<tr>
<td>Multiple Choice Standard</td>
<td>78.9%</td>
<td>71.8%</td>
<td>69.4%</td>
<td>72.7%</td>
<td>69.2%</td>
<td>74.6%</td>
<td>68.9%</td>
<td>65.7%</td>
</tr>
<tr>
<td>Multiple Response</td>
<td>16.0%</td>
<td>22.3%</td>
<td>23.1%</td>
<td>25.5%</td>
<td>15.2%</td>
<td>13.5%</td>
<td>17.7%</td>
<td>16.3%</td>
</tr>
<tr>
<td>Ordered Response</td>
<td>0.7%</td>
<td>1.8%</td>
<td>1.4%</td>
<td>0.0%</td>
<td>4.7%</td>
<td>3.6%</td>
<td>5.1%</td>
<td>2.9%</td>
</tr>
</tbody>
</table>

*Note: 13 items were not coded with a specific topic, and are not included in the table.*

4.2 Responses per Item

The number of responses per item was similar across all item formats. With a minimum of 401 responses per item, the item parameter estimates are stable. Items
that received less than 400 responses were dropped from the study (4 multiple-choice standard items were dropped) to ensure stable item parameter estimates. The overall mean number of responses was 526.7, the median was 515, the minimum was 401, the maximum was 714, and the standard deviation was 63.0. The skewness of the distribution was 0.55 and the kurtosis was -0.64. A D’Agostino-Pearson Omnibus test for normality produced a value of 0.71.

To check for bias in the random distribution of items to test-takers, the mean number of responses per item was compared by item format. An ANOVA using the number of responses as the dependent variable, and item format as the independent variable, produces an F value of 4.49 with a p-value of 0.0002. Because the p-value was below the alpha of 0.01 used for this study, a pairwise comparison was conducted using the Tukey-Kramer method. The results show no significant differences between response counts for any of the pairwise comparisons. The significant F statistic may be an artifact of the large sample size used for this study. Figure 1 below is a box plot showing the distribution of responses by item format.
Figure 1: Distribution of response counts by item format.

4.3 Item Difficulty

The overall item difficulty mean is -0.20, the median is -0.14, the minimum is -6.22, the maximum is 6.02, and the standard deviation is 1.63. The skewness of the distribution is -0.16 and the kurtosis is 0.55. A Komogorov-Smirnov test for normality produces a p-value of <0.01 so a QQ plot was created to assess normality.
The QQ plot shows a generally normal distribution with the exception of a few extreme outliers. The ANOVA procedure is robust to deviations from the normality assumption so an ANOVA procedure, adjusted for the unbalance sample sizes, was used for the analysis.

An ANOVA using item difficulty in logits as the dependent variable, and item format as the independent variable, produces an F value of 107.84 with a p-value of <0.0001. Because the p-value was below the alpha of 0.01 used for this study, a pairwise comparison was conducted using the Tukey-Kramer method. The results show
significant differences in item difficulty between multiple response and all other item formats, and between ordered response and all other item formats except multiple-choice priority and calculation items.

Table 5 below provides the mean and standard deviation of the item difficulty by item format. Table 6 shows the results of the ANOVA. Table 7 provides P-values for the pairwise comparisons, and Figure 3 is a box plot showing the distribution of the item difficulty by item format.

Table 5: Descriptive statistics for item difficulty

<table>
<thead>
<tr>
<th>Item Format</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calculation</td>
<td>-0.38</td>
<td>1.76</td>
</tr>
<tr>
<td>Multiple Choice Follow-up</td>
<td>-0.73</td>
<td>1.39</td>
</tr>
<tr>
<td>Multiple Choice Graphic/Exhibit</td>
<td>-0.44</td>
<td>1.38</td>
</tr>
<tr>
<td>Multiple Choice Priority</td>
<td>-0.24</td>
<td>1.18</td>
</tr>
<tr>
<td>Multiple Choice Standard</td>
<td>-0.49</td>
<td>1.50</td>
</tr>
<tr>
<td>Multiple Response</td>
<td>0.96</td>
<td>1.66</td>
</tr>
<tr>
<td>Ordered Response</td>
<td>0.39</td>
<td>1.75</td>
</tr>
<tr>
<td>Overall</td>
<td>-0.20</td>
<td>1.63</td>
</tr>
</tbody>
</table>
### Table 6: Item difficulty ANOVA results

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>6</td>
<td>1,508.93</td>
<td>251.48</td>
<td>107.84</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Error</td>
<td>4,699</td>
<td>10,958.16</td>
<td>2.33</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>4,705</td>
<td>12,467.09</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R-Square</td>
<td></td>
<td></td>
<td></td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td>Coeff Var</td>
<td></td>
<td>-749.68</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Root MSE</td>
<td></td>
<td>1.53</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td>-0.20</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 7: P-values for item difficulty comparisons

<table>
<thead>
<tr>
<th></th>
<th>MCF</th>
<th>MCG</th>
<th>MCP</th>
<th>MCS</th>
<th>MR</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calculation</td>
<td>0.8290</td>
<td>1.0000</td>
<td>0.9985</td>
<td>0.9969</td>
<td>&lt;.0001</td>
<td>0.0165</td>
</tr>
<tr>
<td>Multiple Choice</td>
<td></td>
<td></td>
<td>0.8005</td>
<td>0.4606</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Follow-up (MCF)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple Choice</td>
<td></td>
<td></td>
<td>0.9688</td>
<td>0.9989</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Graphic/Exhibit (MCG)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple Choice</td>
<td></td>
<td></td>
<td>0.8137</td>
<td>&lt;.0001</td>
<td>0.0773</td>
<td></td>
</tr>
<tr>
<td>Priority (MCP)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple Choice</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Standard (MCS)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple Response (MR)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.0014</td>
</tr>
</tbody>
</table>
4.4 Point Measure Correlation

The overall point measure correlation is 0.09, the median is 0.09, the minimum is -0.18, the maximum is 0.35, and the standard deviation is 0.07. The skewness of the distribution is 0.13 and the kurtosis is -0.01. The Kolmogorov-Smirnov test produces a p-value of 0.15. An ANOVA procedure, adjusted for the unbalanced sample sizes, was used to analyze the variable.

An ANOVA using the point measure correlation as the dependent variable, and item format as the independent variable, produces an F value of 10.63 with a p-value of <0.0001. Because the p-value was below the alpha of 0.01 used for this study, a pairwise comparison was conducted using the Tukey-Kramer method. The results show
significant differences in point measure correlation between calculation and multiple response, ordered response, multiple choice standard and multiple choice follow-up. The multiple response and ordered response item formats both had significantly different point estimates for the point measure correlation than both the multiple choice graphic/exhibit and multiple choice standard item formats.

Table 8 below provides the mean and standard deviation of the point measure correlation by item format. Table 9 provides the results of the ANOVA. Table 10 provides P-values for the pairwise comparisons, and Figure 4 is a box plot showing the distribution of the point measure correlation by item format.

Table 8: Descriptive statistics for point measure correlations

<table>
<thead>
<tr>
<th>Item Format</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calculation</td>
<td>0.13</td>
<td>0.08</td>
</tr>
<tr>
<td>Multiple Choice Follow-up</td>
<td>0.08</td>
<td>0.06</td>
</tr>
<tr>
<td>Multiple Choice Graphic/Exhibit</td>
<td>0.11</td>
<td>0.07</td>
</tr>
<tr>
<td>Multiple Choice Priority</td>
<td>0.10</td>
<td>0.06</td>
</tr>
<tr>
<td>Multiple Choice Standard</td>
<td>0.09</td>
<td>0.07</td>
</tr>
<tr>
<td>Multiple Response</td>
<td>0.08</td>
<td>0.07</td>
</tr>
<tr>
<td>Ordered Response</td>
<td>0.07</td>
<td>0.07</td>
</tr>
<tr>
<td><strong>Overall</strong></td>
<td>0.09</td>
<td>0.07</td>
</tr>
</tbody>
</table>
Table 9: Point measure correlation ANOVA results

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>6</td>
<td>0.34</td>
<td>0.06</td>
<td>10.63</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Error</td>
<td>4,699</td>
<td>25.13</td>
<td>0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>4,705</td>
<td>25.47</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R-Square</td>
<td></td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coeff Var</td>
<td></td>
<td>82.79</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Root MSE</td>
<td></td>
<td>0.07</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td>0.09</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 10: P-values for point measure correlation comparisons

<table>
<thead>
<tr>
<th></th>
<th>MCF</th>
<th>MCG</th>
<th>MCP</th>
<th>MCS</th>
<th>MR</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calculation</td>
<td>0.0011</td>
<td>0.1444</td>
<td>0.0838</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Multiple Choice</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Follow-up (MCF)</td>
<td>0.2225</td>
<td>0.8715</td>
<td>0.9894</td>
<td>0.9996</td>
<td>0.6135</td>
<td></td>
</tr>
<tr>
<td>Multiple Choice</td>
<td>0.9862</td>
<td>0.0316</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Graphic/Exhibit (MCG)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple Choice</td>
<td>0.9531</td>
<td>0.3689</td>
<td>0.0355</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Priority (MCP)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple Choice</td>
<td>0.0073</td>
<td>0.0036</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard (MCS)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple Response (MR)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.3602</td>
</tr>
</tbody>
</table>
The overall unweighted standardized mean square fit value is 1.11, the median is 0.87, the minimum is -3.23, the maximum is 8.00, and the standard deviation is 1.27.

The skewness of the distribution is 1.02 and the kurtosis is 1.63. The Kolmogorov-Smirnov test produces a p-value of 0.01, so a QQ plot was created to assess normality (See Figure 5 below).
Figure 5: QQ plot for unweighted standardized mean square fit value.

The QQ plots indicate some deviation from normality, so both an unbalanced ANOVA and a Kruskal Wallis procedure were used for the analysis. The unbalanced ANOVA, using the unweighted standardized mean square fit value as the dependent variable and item format as the independent variable, produces an F value of 7.15 with a p-value of <0.0001. Because the p-value was below the alpha of 0.01 used for this study, a pairwise comparison was conducted using the Tukey-Kramer method. The results show significant differences in unweighted standardized mean square fit values between calculation and all other item formats. The Kruskal-Wallis test produced a p-
value of 0.0001. The post-hoc pairwise comparisons, using the Bonferroni adjustment, showed the same pairwise differences as the ANOVA.

Table 11 below provides the mean and standard deviation of the unweighted standardized mean square fit value by item format. Table 12 provides P-values for the pairwise comparisons, and Figure 6 is a box plot showing the distribution of the unweighted standardized mean square fit value by item format.

<table>
<thead>
<tr>
<th>Item Format</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calculation</td>
<td>0.22</td>
<td>0.71</td>
</tr>
<tr>
<td>Multiple Choice Follow-up</td>
<td>1.08</td>
<td>1.14</td>
</tr>
<tr>
<td>Multiple Choice Graphic/Exhibit</td>
<td>0.95</td>
<td>1.29</td>
</tr>
<tr>
<td>Multiple Choice Priority</td>
<td>1.30</td>
<td>1.18</td>
</tr>
<tr>
<td>Multiple Choice Standard</td>
<td>1.12</td>
<td>1.30</td>
</tr>
<tr>
<td>Multiple Response</td>
<td>1.13</td>
<td>1.13</td>
</tr>
<tr>
<td>Ordered Response</td>
<td>1.37</td>
<td>1.25</td>
</tr>
</tbody>
</table>

Overall | 1.11 | 1.27 |
Table 12: P-values for unweighted standardized mean square fit value comparisons

<table>
<thead>
<tr>
<th></th>
<th>MCF</th>
<th>MCG</th>
<th>MCP</th>
<th>MCS</th>
<th>MR</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calculation</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Multiple Choice</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Follow-up (MCF)</td>
<td>0.2508</td>
<td>0.1556</td>
<td>0.9243</td>
<td>0.6014</td>
<td>0.1094</td>
<td></td>
</tr>
<tr>
<td>Multiple Choice</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Graphic/Exhibit (MCG)</td>
<td>0.0100</td>
<td>0.0470</td>
<td>0.0094</td>
<td>0.0016</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple Choice</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Priority (MCP)</td>
<td>0.0673</td>
<td>0.1523</td>
<td>0.8596</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple Choice</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard (MCS)</td>
<td>0.1300</td>
<td>0.0125</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple Response (MR)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.0785</td>
</tr>
</tbody>
</table>
The overall weighted standardized mean square fit value is 0.90, the median is 0.46, the minimum is -3.30, the maximum is 7.85, and the standard deviation is 1.19. The skewness of the distribution is 1.50 and the kurtosis is 2.79. The Kolmogorov-Smirnov test produces a p-value of 0.01, so a QQ plot was created to assess normality (See Figure 7 below).
Figure 7: QQ plot for weighted standardized mean square fit value.

The QQ plots indicate some deviation from normality so both an unbalanced ANOVA and a Kruskal-Wallis procedure were used for the analysis. An unbalanced ANOVA using the weighted standardized mean square fit value as the dependent variable, and item format as the independent variable, produces an F value of 5.67 with a p-value of <0.0001. Because the p-value was below the alpha of 0.01 used for this study, a pairwise comparison was conducted using the Tukey-Kramer method. The results show significant differences in the Weighted Standardized Mean Square Fit Value between calculation and all other item formats. The Kruskal-Wallis test produced a p-
value of 0.0001. The post-hoc pairwise comparisons, using the Bonferroni adjustment, showed the same pairwise differences as the ANOVA.

Table 13 below provides the mean and standard deviation of the weighted standardized mean square fit value by item format. Table 14 provides P-values for the pairwise comparisons, and Figure 8 is a box plot showing the distribution of the weighted standardized mean square fit value by item format.

<table>
<thead>
<tr>
<th>Item Format</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calculation</td>
<td>0.12</td>
<td>0.43</td>
</tr>
<tr>
<td>Multiple Choice Follow-up</td>
<td>0.83</td>
<td>1.14</td>
</tr>
<tr>
<td>Multiple Choice Graphic/Exhibit</td>
<td>0.81</td>
<td>1.18</td>
</tr>
<tr>
<td>Multiple Choice Priority</td>
<td>1.11</td>
<td>1.15</td>
</tr>
<tr>
<td>Multiple Choice Standard</td>
<td>0.92</td>
<td>1.23</td>
</tr>
<tr>
<td>Multiple Response</td>
<td>0.88</td>
<td>1.05</td>
</tr>
<tr>
<td>Ordered Response</td>
<td>1.03</td>
<td>1.29</td>
</tr>
<tr>
<td>Overall</td>
<td>0.90</td>
<td>1.19</td>
</tr>
</tbody>
</table>
Table 14: P-values for weighted standardized mean square fit value comparisons

<table>
<thead>
<tr>
<th></th>
<th>MCF</th>
<th>MCG</th>
<th>MCP</th>
<th>MCS</th>
<th>MR</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calculation</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Multiple Choice</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Follow-up (MCF)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple Choice</td>
<td>0.4386</td>
<td>0.1091</td>
<td>0.8884</td>
<td>0.6128</td>
<td>0.4255</td>
<td></td>
</tr>
<tr>
<td>Graphic/Exhibit (MCG)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple Choice</td>
<td>0.0205</td>
<td>0.1095</td>
<td>0.0561</td>
<td>0.0518</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Priority (MCP)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple Choice</td>
<td>0.0598</td>
<td>0.0923</td>
<td>0.3622</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard (MCS)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple Response (MR)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.4670</td>
</tr>
</tbody>
</table>
Figure 8: Weighted standardized mean square fit value by item format

4.7 Response Time

The overall mean response time in seconds is 65.10, the median is 62.06, the minimum is 16.70, the maximum is 364.47, and the standard deviation is 25.03. The skewness of the distribution is 2.72 and the kurtosis is 15.33. The Kolmogorov-Smirnov test produced a p-value of <0.01, so a QQ plot was created to assess normality (See Figure 9 below).
The QQ plots indicate some deviation from normality so both an unbalanced ANOVA and a Kruskal-Wallis procedure were used for the analysis. An unbalanced ANOVA using the response time as the dependent variable, and item format as the independent variable, produces an F value of 631.95 with a p-value of <0.0001. Because the p-value was below the alpha of 0.01 used for this study, a pairwise comparison was conducted using the Tukey-Kramer method. The results show significant differences in response time for all pairs except multiple choice graphic/exhibit and multiple response, multiple choice standard and multiple choice priority, multiple choice standard and

Figure 9: QQ plot for response time.
multiple choice follow-up and multiple choice follow-up and multiple choice priority. The Kruskal-Wallis test produced a p-value of 0.0001. The post-hoc pairwise comparisons, using the Bonferroni adjustment, showed slightly different results than the ANOVA so the results from the post-hoc Wilcoxon comparisons were used for the analysis.

Table 15 below provides the mean and standard deviation of the response time by item format. Table 16 provides P-values for the pairwise comparisons, and Figure 10 is a box plot showing the distribution of the response time by item format.

<table>
<thead>
<tr>
<th>Item Format</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calculation</td>
<td>188.10</td>
<td>42.83</td>
</tr>
<tr>
<td>Multiple Choice Follow-up</td>
<td>53.91</td>
<td>13.44</td>
</tr>
<tr>
<td>Multiple Choice Graphic/Exhibit</td>
<td>73.07</td>
<td>34.99</td>
</tr>
<tr>
<td>Multiple Choice Priority</td>
<td>58.02</td>
<td>14.35</td>
</tr>
<tr>
<td>Multiple Choice Standard</td>
<td>59.52</td>
<td>16.41</td>
</tr>
<tr>
<td>Multiple Response</td>
<td>71.82</td>
<td>17.22</td>
</tr>
<tr>
<td>Ordered Response</td>
<td>102.32</td>
<td>29.17</td>
</tr>
<tr>
<td>Overall</td>
<td>65.10</td>
<td>25.03</td>
</tr>
</tbody>
</table>
Table 16: P-values for response time comparisons

<table>
<thead>
<tr>
<th></th>
<th>MCF</th>
<th>MCG</th>
<th>MCP</th>
<th>MCS</th>
<th>MR</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calculation</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Multiple Choice</td>
<td></td>
<td>&lt;.0001</td>
<td>0.0462</td>
<td>0.0012</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Follow-up (MCF)</td>
<td></td>
<td></td>
<td>0.0180</td>
<td>0.0002</td>
<td>0.0033</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Multiple Choice</td>
<td></td>
<td></td>
<td></td>
<td>0.5901</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Priority (MCP)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Standard (MCS)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Multiple Response (MR)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>
Figure 10: Response time by item format.
CHAPTER 5. DISCUSSION

5.1 Introduction

The purpose of this study was to detect differences in item parameter estimates due to item formats from a professional licensure examination program. The items did indeed have some parameters that were not different statistically, however, there are some differences that can inform the way items are used, or indicate that an item format should be studied further before it is accepted as a general tool for test development. In particular, item discrimination and response time showed significant differences between item formats.

5.2 Sample Size

The samples collected for this study are unequal. The largest sample (3,330 for multiple-choice standard items) is 52 times larger than the smallest sample (64 for calculation items). The sample of 64, however, is sufficient to detect differences with a reasonable degree of confidence. The analysis uses an alpha level of 0.01 with the Tukey-Kramer adjustment for familywise error for the ANOVAs, and a Bonferroni adjustment for the Kruskal-Wallis procedures. A post-hoc power analysis, using the
smallest sample size (64), produces a lower limit of the beta level of 0.97 for effect sizes of 1 standard deviation or more.

While the sample sizes are unequal, the relatively high alpha and beta levels for this study provide evidence that any differences uncovered are true differences, and are not statistical artifacts. This is particularly true for the multiple choice and multiple response item formats although the sample sizes for calculation and ordered response item formats are sufficiently large enough to draw reliable inferences.

5.3 Item Difficulty

The significant differences in item difficulty were for multiple response items and for ordered response items. The higher mean difficulty of multiple response items is likely an artifact of the dichotomous scoring used for the test. There are two reasons for this. First, examinees are informed that two or more options are correct but not exactly how many are correct, and the examinee must identify all options correctly (as correct or incorrect) in order to receive credit for the item. This scoring structure creates a much different probability for guessing a correct response compared to a standard multiple-choice item.

Assuming a completely random guess, a standard four-option multiple-choice item has a guessing probability of approximately 0.25. Multiple response items, as presented and scored on this examination, have a guessing probability of 0.04 with five
options and 0.02 with six options. \[ P(\text{correct response}) = \frac{1}{\sum_{x} \frac{n!}{n!(n-i)!}} \text{ where } x = \text{number of options}. \]

Secondly, dichotomously scored multiple response items by nature confound the measure of knowledge. For example, an examinee may know three of the four correct answers for a particular multiple response item and receive no credit for the item. Contrast this to a set of four multiple-choice items where the candidate would get three out of four items correct. The probability of guessing a correct response to three out of four multiple-choice items (0.05) is similar to guessing the correct response to a single five-option multiple response item (0.04).

The difference in difficulty for ordered response items may be due, in part, to the reduced probability of randomly guessing the correct response. A five-option ordered response item has a probability of randomly guessing a correct response of 0.01 \([1/5!]\) (the factorial, !, is used to calculate the number of possible responses) compared to 0.04 for a five-option multiple response and 0.25 for a four-option multiple-choice item.

5.4 Point Measure Correlation

Calculation items had the best relative point measure correlation estimates of any of the item formats. This conforms to theory in that item formats with lower probabilities of guessing the correct answer should exhibit better discrimination. The more interesting finding is that the multiple response and ordered response item
formats, both of which have lower guessing probabilities compared to multiple choice item formats, had significantly lower point measure correlation estimates in comparison to the multiple-choice standard item format.

This may be due to the dichotomous scoring method used for both item formats. More information and perhaps better discrimination could be obtained by using a polytomous scoring method. The significant differences could also be interpreted as evidence that these formats are not distinguishing as well as multiple-choice formats, and may not be contributing as much to the reliability of examination scores.

5.5 Unweighted and Weighted Standardized Mean Fit Values

Unweighted and weighted standardized mean square fit values have similar results and interpretations. The ideal value for fit statistics (unweighted or weighted standardized mean square fit value) under the Rasch model is zero. Therefore, deviations from zero, either positive or negative, are of interest. For all item formats, the mean was above zero. The only significant difference was between calculation items and all other item formats. In this study, calculation items show the best model fit.

Multiple response and ordered response also have a low probability of randomly guessing the correct response when compared to standard multiple-choice items and should, in theory, show better item fit. However, the fit for multiple response and ordered response formats are not empirically distinct in this sample in comparison to multiple-choice items (all varieties).
The lack of an empirical difference between the fit of multiple response and ordered response items in comparison to multiple-choice items may be due to other issues that affect model fit. The slightly lower, but statistically significant, point measure correlation for both multiple response and ordered response compared to multiple choice indicates that these two item formats may not be separating high-performing candidates from low-performing candidates as well as multiple-choice items.

In the case of multiple response items, the fit might be improved by adopting a polytomous scoring system. If polytomous scoring works properly, the items could have a very favorable information per unit time ratio. For ordered response items, there may be a problem inherent in the item format. While in theory, ordered response items could be scored polytomously, setting up a rubric is difficult and reaching consensus on item order is problematic. On a practical level, it may not be worth the time and effort required to create these ordered and multiple response item types if they do not add to ability measurement is a significant way.

5.6 Response Time

The largest variations between item formats were in response time. Multiple-choice items of all varieties, except graphic/exhibit items, had average response times between 53 and 60 seconds. They also had similar standard deviations of between 13 and 16 seconds. Multiple choice standard, multiple choice follow-up, and multiple choice priority have statistically shown non-significant differences in response time.
Multiple choice graphic/exhibit items have a longer mean response time of 73 seconds, and a larger standard deviation of 35 seconds. This is likely due to the larger amount of material to process in graphic/exhibit items, and possibly due to time required to navigate between screens for these particular items. Given the significantly longer response time required for this format item, it should only be used when a graphic or exhibit is essential to testing the content. Since this item cannot be scored polytomously, its information per unit time ratio is lower than all the other formats except for calculations.

The mean response times for multiple choice graphic/exhibit and multiple response items are not significant, however, the standard deviations in response time were 35 seconds for multiple choice graphic/exhibit and 17 seconds for multiple response items. Ordered response items had significantly different response times (102 seconds) than all other items, and had a relatively large variance in response time (standard deviation = 29 seconds). The largest difference lies with calculation items, which have a mean response time of 188 seconds with a standard deviation of 43 seconds.

The extended response time for ordered response is quite interesting because the number of options is no greater than for the multiple response items; nonetheless, the examinees took 40% longer to respond to the ordered response items. A similar relationship exists between the standard deviation estimates with ordered response items showing greater variability in response time. Since all of the options are present and visible on one screen, the greater response time may be due to examinees being
uncertain about their responses, or they have difficulty with the mechanics of responding to the items.

Ordered response items are presented on a single screen with all available options visible. The options are left justified in a column on the left hand side of the screen, and examinees are instructed to click and drag items from the column on the left into the correct order in the column on the right. Examinees may move the options in any order, and may rearrange options in the right column. It is possible that this response method is at least partially responsible for the increased response time, however; examinees for this examination are generally computer literate, and have had an opportunity to work with all of the item formats on the examination before taking it. Some students may know the OR items right away, and others may struggle with various sequences before finalizing their response.

The largest difference in response time is for the calculation items. In general, this is not surprising since it has been previously established that it is more difficult for examinees to construct a correct answer than to recognize one from a list. Another factor which could account for the additional response time is the time needed by examinees to determine the correct formula to apply and to complete the calculation. The sizeable difference in response time has implications for testing time limits, as well as the examinee’s experience. Tests which contain larger numbers of calculations or ordered response items could create time pressure on examinees or create a greater fatigue effect, particularly on lengthy examinations. This may be especially true for lower ability candidates since calculation items tended to have lower item difficulties.
5.7 Comparison of Item Formats

The similar results across all five item parameters for all variants of the multiple choice item suggest that including graphics or exhibits, requiring an examinee to prioritize choices or choose the exception in the options all work similarly. This is welcome news from a test development perspective. Since all these formats work equally well, the most appropriate or convenient format for the content can be used.

Calculation items produced good item parameter values, particularly in regard to item fit and the point-measure correlation. However, the relatively long mean response time indicates that this item format cannot be used interchangeably with multiple choice items. This raises two important issues. First, examinees who received relatively high numbers of calculation items may need more time to complete an examination and second, calculations may cause more fatigue for examinees. Both of these issues are manageable on a practical level since this item type only works for a specific type of content. A range of the number of calculation items on an examination could be included as part of the test specifications.

Calculation items test a specific skill within the construct. The relatively easy difficulty of the item format suggests that this is a skill that most candidates have mastered, however, the relative long response time suggests that a different cognitive process is being used to respond to the items. While many different skills are included in the construct, it is helpful to recognize that this particular skill, while not extremely complex, does take more time to think and calculate a response.
The response time for calculation items could be particularly troublesome. The mean difficulty for calculation items is low. On an adaptive test, lower ability candidates would have a greater probability of receiving these items. This would increase the overall amount of time that these candidates spend testing, which could increase test fatigue and cause candidates to run out of time. It might be wise to limit the number of these items that can be administered to a test-taker during a testing session.

Multiple response items produced overall acceptable item parameters, and were significantly more difficult than all of the other item types. This is likely due to the dichotomous scoring of the items. Polytomous scoring could potentially make this item type more useful as it could provide more information than dichotomously scored items, and might improve the item discrimination. However, the increased difficulty, particularly given that items fit the model using dichotomous scoring, can be useful when examinations are given as an adaptive examination. The additional difficulty could provide items more closely matched to high-performing candidates, making the adaptive test more efficient.

Ordered response items were slightly more difficult than multiple-choice items, slightly less difficult than multiple response items, and they had the lowest point measure correlation of all the item types. While their model fit was similar to other item formats, the mean response time for ordered response was approximately 60% higher than for multiple choice. Other than the increased response time, no one parameter stands out, but the combination of lower point-measure correlation, increased response
time, and increased difficulty suggests that this item type may have a higher degree of construct irrelevant variance than others may in this study.

A discussion with professional content developers suggests that the difficulty might lie in developing items with five distinct options with a definite order. In many processes, there are at least some steps which have to occur but which are interchangeable. As an example, consider the following sequence for starting an automobile:

1. Put on your seat belt.
2. Place the key in the ignition.
3. Place your foot on the brake.
4. Turn the key.
5. Release the key when the engine starts.

Even assuming a strict interpretation of the law (the seat belt must be put on before the key goes into the ignition), steps two and three could be interchanged with no effect on the outcome of the process. Developing items with no interchangeable steps is particularly difficult for short processes. Leaving out steps as they were learned by the examinee can lead to confusion, and thereby diminish the face validity of the item.

One way to address this issue would be to develop a scoring rubric which grants credit for multiple orders. In the example above, credit could be given for the sequence as demonstrated, or with steps 2 and 3 interchanged. This change would increase the probability of randomly guessing a correct answer slightly (from 0.01 to 0.02 for the example above) but it might improve the discrimination of the item.
This analysis does not conclusively show that ordered response items have a greater amount of construct irrelevant variance but it is indicative. Subject matter expert comments suggest that the content in ordered response items is not notably different than content in other item formats. It is not clear why ordered response items have a greater response time in comparison to other selected response item formats. It is also not clear why ordered response items would be more difficult than multiple choice items given the similar nature of the response choices.

It is worth considering whether the ordered response item format is a good indicator of the understanding of steps in a process. Given that many processes have interchangeable or optional steps, this item format may not accurately reflect a candidate’s knowledge. Given the additional effort required to develop and validate these items, the limited content for which it can be used and the apparent higher level of content irrelevant variance, it might be prudent to limit the use of this item type.

5.8 Limitations

There are two important limitations to this study—the composition of the sample, and the unidimensionality of the construct.

The first limitation is the sample of test-takers. The sample is based on a high-stakes, professional licensure examination involving adult test-takers. The results may not apply to lower stakes examinations or to younger test-takers, particularly where the reading level of the item format plays a greater role in the estimation of item parameters.
Candidates for the licensure pass through a multi-step process before they are allowed to take the examination. Candidates self-select to apply for a professional program, must meet the selection criteria to be admitted to a professional program, must pass a multi-year college curriculum and maintain an appropriate grade point average, must pass clinical tests of their skills, and must meet additional regulatory requirements before being allowed to sit for the licensure examination. This results in a large group of people with very similar professional knowledge, skills, and abilities. Any correlation analysis on such a population is attenuated because of the narrow range. This applies most directly to the point measure correlation used in this study, but also has implications for generalizability of the conclusions about the item formats. This limitation can be overcome in future studies by doing similar research using more divergent groups of test-takers. One option would be to include both newly graduated candidates and experienced clinicians.

The sample also limits the generalizability of these results especially as they apply to other populations. While no effect was found for follow-up multiple-choice items (negative wording), other studies have found an effect for younger populations. Due to differences in reading level, motivation, maturity, and preparation, the results of this study may not apply to other populations.

The unidimensionality of the construct is harder to address because some researchers contend that an item parameter such as item difficulty cannot be compared between item formats even if identical content is used because the underlying true
score scales cannot be shown to be equivalent (Traub, 1993, p. 30). Moreover, all but the simplest constructs will have some level of multidimensionality.

Several arguments have been presented to support the comparison of item parameters within this construct, but it cannot be proved conclusively that the effects observed here are not confounded with multidimensionality. Future studies may be able to address this issue by using simpler constructs, or by creating item sets with matched content.

5.9 Future Research

The results of this study suggest that at least two follow-up research projects require further investigation: 1) Examining the effect of polytomous scoring with the multiple response item format, and 2) examining the effect of a more inclusive scoring rubric for the ordered response item format.

The study of the multiple-choice response format would consist of two parts. The first step would be to take a sample of multiple response items with two or three correct options and three incorrect options. The key feature would be having three incorrect options for each item. The original items would be used to create variants where the stems would remain constant, and one correct option with all the incorrect options would be used to create a multiple-choice item.

For multiple response items with two correct options, this would create an original plus two variants; for multiple response items with three correct options, this would create an original plus three variants. There would be no changes to the content,
only a rearranging of the options. The items would then be given to a random sample of examinees ensuring that only the original or a variant of each item is presented to a single examinee. This would avoid any issues with item independence or item cueing.

The analysis would consist of comparing the item parameters of the original items with the variants, and then determining whether a monotonic relationship exists between the different response levels. If a monotonic relationship is present (polytomous scoring is not useful for measurement if a monotonic relationship does not exist), then several different polytomous scoring methods could be compared.

This would a particularly interesting study because polytomous scoring for multiple response items could produce a much higher test information to response time ratio. This would allow test developers to test a wider range of information, and more quickly reduce the standard error of measurement per unit of candidate time.

Determining whether item parameters would improve for the ordered response item format if a more inclusive scoring rubric was used would also consist of two parts. A sample of ordered response items would need to be evaluated by subject matter experts to determine if any of the steps were interchangeable (if the item has no interchangeable steps, it is not suitable for this study). These steps would need to be identified, and a scoring rubric prepared for each item.

Depending on the sample, several conditions could be examined. For example, there might be two or three interchangeable options (although three interchangeable options significantly changes the probability of guessing a correct response). There also may be two sets of interchangeable steps with a set of options (again, this would create
a significant increase in the probability of randomly guessing a correct response). The items would be given to examinees, and then the response patterns would be analyzed using the original single correct answer rubric and the variant rubric created for the study.

5.10 Conclusion

The item formats included in this study generally produce useable item parameters. Item fit indices show that the item types used in this examination are appropriate for the purposes of the test. The relative balance in the distribution of item fit and item difficulty across the variety of item formats suggests that test developers have a lot of latitude in creating item content for professional licensure examinations.

As has been shown by numerous previous research studies, the multiple-choice item format produces reasonable item parameters with some loss of reliability due to guessing. Adding graphics or exhibits to the item format affects response time slightly but does not inherently affect item difficulty or discrimination. Including content that requires clinical judgment skills (priority type) or the ability to discern an exception (follow-up type) also does not significantly affect item parameters. Multiple-choice item formats have definite limits (for example: inability to test divergent thinking, creativity, or recall of information), however, given the context of these limits, this item format is useful and efficient.

Multiple response items require more response time generally than do MC items, but have item discrimination parameters similar to MC items. This an interesting finding
because multiple response items in this study use a dichotomous scoring model. Using a polytomous scoring model for multiple response items would yield greater information per item and per time unit. Although assignment of partial scoring points to each item would be more time-consuming, the use of polytomous scoring could make testing more efficient, and enable test developers to cover more content and to increase the reliability of a test or reduce the test length. While the multiple response item format suffers from many of the same limits as the multiple-choice item format, the potential for increasing testing efficiency is promising.

The ordered response items in this study produced item parameters that suggest that this item type may be introducing a higher level of construct irrelevant variance than the other item types included in the study. The longer response time, somewhat lower point measure correlation and model fit suggest that this item format may not be as useful or efficient as other options. Additionally, because the variety of tasks that can be tested using OR items are relatively limited, it might be prudent to limit or avoid this type of item format until further research has explored more fully the characteristics of this item type.

Unsurprisingly, calculation items demonstrated the best overall fit to the measurement model. With their low probability of randomly guessing the correct response, calculation items more closely meet the assumptions of the Rasch model used in this study (which assumes a zero guessing parameter). Calculation items were found to be relatively easier but more time-consuming than all other item types evaluated in this study. This may indicate that the calculation item format is less efficient in terms of
response time, but more efficient in terms of reliability and appropriateness for the model. The calculation item format does not have the same limits as selected response formats, and the ability to perform calculations is an essential part of many constructs. The combination of increased response time and lower relative item difficulty poses some challenges for adaptive tests but the item format produces useful parameters and tests content that cannot easily be tested using other item formats.

Based on the data from this study, in terms of item difficulty and discrimination, multiple choice, multiple response and calculation item formats are valuable tools for test development. Ordered response items should be used judiciously when other item formats are not appropriate. In terms of response time, test developers should be aware of the relatively longer response time for calculation items, and limit the use and number of these item types in an operational testing environment.
REFERENCES
REFERENCES


distractors in multiple-choice Items: Deficiencies in usage and an alternative.

*Journal of Educational Measurement, 37*(1), 77-86.

Embretson (ed.). *Measuring psychological constructs: Advances in model-based


Benson, J. & Crocker, L. (1979). The effects of item format and reading ability on
objective test performance: A question of validity. *Educational and Psychological
Measurement, 39*, 381-387.

elementary school children*. Paper presented at the annual meeting of the
American Educational Research Association, Boston, MA.

Erlbaum.


Council on Education.


Yang, S. (2007). *A comparison of unidimensional and multidimensional Rasch models using parameter estimates and fit indices when assumption of unidimensionality is violated*. (Unpublished doctoral dissertation). The Ohio State University, Columbus, OH.

VITA

Gregory M. Applegate
Graduate School, Purdue University

Education
B.A., History, 1991, Indiana University, Fort Wayne, Indiana
M.B.A., Business Administration, 1999, Indiana University, Fort Wayne, Indiana
Ph.D., Educational Psychology, 2014, Purdue University, West Lafayette, Indiana

Research Interests
Item Development
Computerized Adaptive Testing
Automatic Item Generation
Technology Enhanced Items

Experience
Manager, Content Development, May 2012 – present, Pearson VUE, Chicago, IL
Psychometric Intern, Aug. 2010 – May 2012, Pearson VUE, Chicago, IL
Continuing Lecturer, Statistics, Aug 2009 – May 2010, Purdue University, Fort Wayne, IN