2010-12-16

An Evaluation of Multiple Choice Test Questions Deliberately Designed to Include Multiple Correct Answers

Kim Scott Thayn
Brigham Young University - Provo

An Evaluation of Multiple Choice Test Questions Deliberately Designed to Include Multiple Correct Answers

Scott Thayn

A dissertation submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Richard R. Sudweeks
Randall S. Davies
David F. Foster
Joseph A. Olsen
David Wiley

Department of Instructional Psychology & Technology

Brigham Young University

April 2011

ABSTRACT

An Evaluation of Multiple Choice Test Questions Deliberately Designed

to Include Multiple Correct Answers

Scott Thayn

Department of Instructional Psychology and Technology

Doctor of Philosophy

The multiple-choice test question is a popular item format used for tests ranging from classroom assessments to professional licensure exams. The popularity of this format stems from its administration and scoring efficiencies. The most common multiple-choice format consists of a stem that presents a problem to be solved accompanied by a single correct answer and two, three, or four incorrect answers. A well-constructed item using this format can result in a high quality assessment of an examinee's knowledge, skills and abilities. However, for some complex, higher-order knowledge, skills and abilities, a single correct answer is often insufficient. Test developers tend to avoid using multiple correct answers out of a concern about the increased difficulty and lower discrimination of such items. However, by avoiding the use of multiple correct answers, test constructors may inadvertently create validity concerns resulting from incomplete content coverage and construct irrelevant variance. This study explored an alternative way of implementing multiple-choice questions with two or more correct answers by specifying in each question the number of answers examinees should select instead of using the traditional guideline to select all that apply. This study investigated the performance of three operational exams that use a standard multiple-choice format where the examinees are told how many answers they are to select. The collective statistical performance of multiple-choice items that included more than one answer that is keyed as correct was compared with the performance of traditional single-answer, multiple-choice (SA) items within each exam. The results indicate that the multiple-answer, multiple-choice (MA) items evaluated from these three exams performed at least as well as to the single-answer questions within the same exams.

Keywords: multiple-answer, multiple-correct, multiple-choice, number correct, test question

ACKNOWLEDGEMENTS

This dissertation is dedicated to my loving wife, Kim, without whom, I would have remained ABD for the rest of my life. Thanks to her love, encouragement and support, this dissertation is only one of many things in my life that I would have never accomplished. She has been my inspiration and my source of strength. When I grow up, I want to be just like her.

I would also like to express my profound gratitude to Dr. Richard Sudweeks, a great mentor and scholar. I consider it a great honor to have been able to work and learn from this great man. His guidance and wisdom have been invaluable throughout my years at BYU, but his involvement on this project truly demonstrated his limitless patience and willingness to share of his time and knowledge.

Finally, I would like to thank my family for their sacrifice and support throughout this project and my many years in school. My greatest hope is that the legacy I leave my children is the desire to learn and the drive to accomplish all of their goals.

Table of Contents

Chapter 1: Introduction 1

Background Information 2

Challenges Faced With Certification and Professional Licensure Exams 4

Inadequacy of Traditional Single-Response, Multiple-Choice Items 4

Item Types Built to Address the Need for Multiple Answers 5

Research Questions 12

Chapter 2: Literature Review 14

Single Correct Answer Recommendation 14

Type K Items 16

Multiple True-False Items 17

Multiple Answer Items 20

Summary 21

Chapter 3: Method 23

The Selection of Exams Used in This Study 23

Design 25

Chapter 4: Results 34

Content Validity of MA Items 34

Acceptance Rates of SA and MA Items 36

Average Statistical Characteristics of SA and MA Items 39

Partial Credit Versus Dichotomous Scoring of MA Items 45

Chapter 5: Discussion 53

Overall Summary and Reflection 53

Interpretation of Findings for Each Research Question 55
Limitations 61
Conclusions 62
Recommendations 62
References 65
Appendix A: Item Classical Statistics for Exam 1 69
Appendix B: Item Classical Statistics for Exam 2 72
Appendix C: Item Classical Statistics for Exam 3 75
Appendix D: Item Rasch Statistics for Exam 1 80
Appendix E: Item Rasch Statistics for Exam 2 83
Appendix F: Item Rasch Statistics for Exam 3 86
Appendix G: SME Ratings for Exam 1 91
Appendix H: SME Ratings for Exam 2 94
Appendix I: SME Ratings for Exam 3 96

List of Tables

Table 1: Random Guessing Probabilities for Multiple-Choice Items....18

Table 2: Examinee Test Records Omitted From This Study....25

Table 3: Statistical Standards....29

Table 4: Summary of Survey Results....35

Table 5: Percentage of SA and MA Items That Satisfy Minimal and Preferred Difficulty Standards....37

Table 6: Comparison of the Item Discrimination Acceptance Percentages....38

Table 7: Comparison of the Item Reliability Acceptance Percentages....38

Table 8: Average Item Analysis Statistics by Exam and Item Type....40

Table 9: Item Statistics for the Dichotomous and Partial Credit Rasch Model....46

Table 10: Reliability Estimates for Dichotomous and Partial Credit Scoring Models....47

List of Figures

Figure 1. Example of a Type K item 6

Figure 2. Example of a Multiple Select Multiple-Choice item 9

Figure 3. Example of a Multiple Select Multiple-Choice Item with Cueing 11

Figure 4. Screen shot of review tool. 28

Figure 5. Average item information function for Exam 1 using only Rasch 42

Figure 6. Average item information function for Exam 2 using only Rasch 42

Figure 7. Average item information function for Exam 3 using only Rasch 43

Figure 8. Average item information function for Exam 1 using two parameters..... 43

Figure 9. Average item information function for Exam 2 using two parameters..... 44

Figure 10. Average item information function for Exam 3 using two parameters..... 44

Figure 11. Test information by form and scoring model for Exam 1 using only Rasch .. 48

Figure 12. Test information by form and scoring model for Exam 2 using only Rasch .. 48

Figure 13. Test information by form and scoring model for Exam 3 using only Rasch .. 49

Figure 14. Test information by form and scoring model for Exam 1 using two parameters 50

Figure 15. Test information by form and scoring model for Exam 2 using two parameters 50

Figure 16. Test information by form and scoring model for Exam 3 using two parameters 51

Figure 17. Score distributions for the dichotomous and Partial Credit models for Exam 1 51

Figure 18. Score distributions for the dichotomous and Partial Credit models for Exam 2 52

Figure 19. Score distributions for the dichotomous and Partial Credit models for Exam 3 52

Chapter 1: Introduction

Testing programs frequently need to measure knowledge, skills and abilities (KSAs) that cannot be adequately measured with traditional multiple-choice items. They often need to measure complex, multi-faceted content that cannot be reasonably constructed into a single-answer, multiple-choice test question. What is needed is an item type that can effectively measure this content while maintaining the desirable characteristics of traditional single-answer, multiple-choice items, such as adaptability for computer administration and automated scoring.

The two primary measures of quality in an exam are validity and reliability. *Validity* addresses whether or not the exam measures what it purports to measure, and *reliability* addresses the consistency of resulting test scores. If the intent of the exam is to measure multi-dimensional skills but the items focus on single-dimensional skills, then the validity of the inferences about the level of an examinee's skills is adversely affected. Additionally, if the structure of the item helps clue the examinee to the correct answer, then validity is also negatively impacted.

The purpose of this research is to find the most efficient and effective manner to measure complex content that can be delivered by most computer delivery software programs. Traditional multiple-choice items are not able to address this need without grouping all parts of the answer into a single option. Not only would this potentially clue the examinee to the correct answer, but it would result in a very inefficient means of testing because each incorrect response should be parallel in structure and content to the correct answer. If the correct answer lists three components or actions, then each distractor should also list thee components or actions. However, to avoid overlapping options, each distractor would need to include discrete information that collectively results in a plausible, yet incorrect answer. Distractors are already

the most difficult part of an item to write, but adding this level of complexity to the requirements of developing plausible distractors is an unnecessary burden on the item writers.

The reasons for the inadequacy of traditional multiple-choice items will be further explored below. However, because of the inadequacy of multiple-choice items, test sponsors have explored the use of various modified item types that allow an examinee to identify more than one correct answer while only requiring a single response.

The premise of this research is that the best solution to measure this content with computer-administered exams is to use the format of the traditional multiple-choice item, but to allow more than one correct option to be included and then specify in the stem how many options examinees are expected to select. While some authors have expressed some concerns in the past about multiple correct answer multiple-choice items (Haladyna & Downing, 1989; Burton, Sudweeks, Merrill, & Woods, 1991), the objective of this research is to show that the concerns that have been expressed are addressed by controlling the administration of the items through the computer delivery of the exams and by instructing the examinee how many options they are to select.

Background Information

Testing objectives are used as a way to identify the specific content that the test sponsor (owner) wants to measure for a given KSA. Testing objectives are written statements of action that identify what is to be tested and what the test sponsor is willing to accept as evidence that the examinee has sufficiently mastered that content.

There has long been a need to assess objectives that require more than a single response (Albanese, 1993; Willson, 1982). Willson (1982) referred to the practice of restricting each item to a single correct answer as a stifling limitation, noting that it is possible to create good

multiple-choice items where several or all options are correct. The National Board of Medical Examiners (NBME) and other examinations in the healthcare professions have incorporated various items types to measure complex, multi-faceted content by requiring examinees to identify more than one component in their responses (Frisbie, 1992; Albanese, 1993). Other professional licensure and certification programs have also used various multiple answer item types to attempt to measure this type of content (Haladyna, 1992). The development of these various item types indicates a need to assess content that goes beyond what a traditional single response item type is able to measure.

For example, suppose an objective on an art exam was to have a student identify which two colors will combine to create the desired color outcome. This objective requires an examinee to identify two separate pieces of information, but it is the knowledge of both of these pieces of information (the two colors) together that either meets or does not meet the single objective. The goal is to measure a student's knowledge of this objective as efficiently and as accurately as possible. While this is a rather simplistic example, this same challenge of measuring content that requires more than a single answer is faced continually by many testing sponsors, especially when high-level knowledge, skills and abilities are being assessed. In addition, it is the desire of most test sponsors to keep the exams as efficient as possible in order to minimize the development and delivery costs for the exam. The cost to develop and to deliver these exams is impacted by the complexity and number of items that are on an exam. Therefore, if test sponsors can measure the same complex content in a much simpler format and with fewer questions, not only can they improve the validity and reliability of their assessments, but they can reduce the development and delivery costs as well.

Challenges Faced With Certification and Professional Licensure Exams

Historically exams were delivered via paper and pencil administration. The preferred item type was multiple choice in order to facilitate automated scoring using a scanning device and computer software. Single-answer, multiple-choice items were required by most scanning and scoring programs, so this was often a requirement for these exams. While there are many certification and licensure exams still delivered and scored this way, the solution to effectively measuring complex objectives will be focused on exams that that are computer-administered exams because of the added control and the greater scoring flexibility the computer affords. However, even though computer administration provides theses added benefits, many testing programs that convert from paper/pencil to computer administration continue to use the same traditional multiple-choice items that were used on the previous paper/pencil exams.

Traditional multiple-choice items consist of a stem and three to five answer options from which the person taking the test can select. The options typically include a single correct answer or keyed response and several plausible but incorrect answers, called distractors. These items are most commonly scored dichotomously, with one point awarded if the correct answer is selected and zero points if any of the incorrect answers are selected.

Inadequacy of Traditional Single-Response, Multiple-Choice Items

In order to measure complex content with a traditional single-response item, item writers often either focus the item on only a portion of the objective or create negatively-worded stems and instruct the examinee to select the answer that is NOT correct. For many testing experts, neither of these options is desirable (Cassels & Johnstone, 1984).

By focusing an item on only one of the multiple correct answers, the rest of the objective is either left untested, or other multiple-choice questions have to be written in order to cover the

entire objective. This practice can have a detrimental impact on content validity by either under sampling or oversampling a particular content domain. It is not uncommon for some objectives to only call for one item. If it takes two items just to cover the entire content of an objective that calls for one item, then that content will be underrepresented if only one item is included on the exam and overrepresented if two or more items are included in order to cover the entire objective. In addition, this would be a highly inefficient and repetitive means of testing, since the stem for each of these multiple questions would be repeated for each item.

One alternative already mentioned is to write a negatively-worded stem such that the examinee is directed to select the option that is NOT correct. With this type of item, an objective with multiple components can be measured; however, the correct answer is what is incorrect about the focus of the stem, which can be a source of confusion and misinterpretation by examinees. Research has shown negatively worded stems to be less discriminating because of the higher tendency to misread or misunderstand the question (Haladyna & Downing, 1989; Cassells & Johnstone, 1984).

Item Types Built to Address the Need for Multiple Answers

Some of the item types that have been developed to address the need to allow more than a single correct answer include multiple true-false (Cronbach, 1941), complex multiple-choice (Albanese, 1982), and muliple-response, multiple-choice (Frary, 1989) items. The latter have been referred to in the literature by various names including (a) multiple-response, multiple-choice, (b) multiple-correct, multiple-choice, and (c) multiple-answer, multiple-choice items. Hereafter, I will refer to items of this type as multiple-answer, multiple-choice items or simply MA items. Complex multiple-choice items, hereafter referred to as CMC items, were developed

to allow for multiple answers to be identified with a single selection. Traditional multiple-choice items with a single correct answer will be referred to as SA items.

Complex multiple-choice items. One such CMC item is the Type-K item (see Figure 1), which was designed to address the combined desire to (a) measure complex objectives that require multiple discrete answers and (b) have the examinee select a single correct answer to facilitate automated scoring. The single answer that the examinee selects, referred to as a secondary option, is comprised of pre-determined combinations of multiple answers, called primary options.

Which item properties are used to calculate item reliability?

I. Item Content
II. Item Kurtosis
III. Item Variance
IV. Item Discrimination

A. I & IV
B. II & III
C. I & IV
D. III & IV

Figure 1. Example of a Type K item

In this example, the examinee must identify that both item discrimination and item variance are used to calculate item reliability, but the two components are combined in discrete pairings to allow the examinee to select the proper combination. This type of item allows for the continued use of computerized scoring programs that only allow one correct answer per item. Haladyna and Downing (1992) cite several reasons for advising against the use of this format, including (a) lower discrimination, (b) higher difficulty, and (c) the effect of partial knowledge which may allow a test-wise examinee to eliminate possible combinations to correctly answer

the question. Other researchers, such as Case and Downing (1989) and Albanese (1982) also found CMC items to have lower reliability than traditional MC items and MTF items.

Multiple True/False items. One item type that has shown promising performance characteristics is the multiple-true-false item, where each option in the item is scored independently, as if it were its own individual item.

While this item type has generally performed better than Type K items, its use presents several challenges in the context of certification and professional licensure testing. The first and most formidable is the limitation of the major computerized test delivery vendors to administer and score this type of item. Another challenge is that by scoring the options as though they were individual true false items, the examinee is essentially being given partial credit for partial knowledge. However, these complex objectives are not merely combinations of multiple mini-objectives that stand on their own, but rather they contain multiple components that must be performed or considered in combination with each other to determine if an examinee is minimally competent. These objectives are defined by the fact that in order to be considered minimally competent in that content area, an examinee must be able to perform the entire objective. By giving partial credit, an examinee could amass enough points through partial knowledge of several content areas as to impact their pass/fail outcome.

These certification and professional licensure exams are all used as a basis for classifying examinees into two categories: passed or failed. This means that a cutoff score has previously been determined for classifying the examinees. Those who earn a total score that equals or exceeds the cutoff are classified in the passed category. On the other hand, examinees' whose score is less than the cut-score are classified as having failed the exam. In order to maximize the reliability and validity of the pass/fail decision, it is desirable to maximize the discrimination

power of each of the items by focusing the difficulty level of each of the items right at the ability level of a minimally-competent examinee. Even though the resulting items never end up having exactly the same difficulty level, they will be grouped in the ability range of the pass/fail decision. The difficulty of each correct answer will not be equivalent, so each scoring opportunity will not be as focused on the decision-making point as the entire item would be. If there are a large number of MA items with a very easy correct answer, the total scores will be inflated for low ability examinees.

The content of a given item is defined by the testing objective which defines which part of the targeted KSA the test sponsor expects the examinee to demonstrate. Each item is expected to meet all requirements set forth in the objective. The individual components of a given complex objective are not necessarily of equal difficulty or relevance, so scoring them as separate questions would weight each component equally by awarding one point per component. But it is not the number of components that should dictate the relative importance of a given objective, but rather the importance and relevance of the objective as a whole.

It is possible to award fractions of a single point to correct answers within a MTF item so that the combined possible points for an item would always equal a single point. However, as Robert Frary (1989) explored various partial credit strategies for scoring multiple answer items, he found little was gained from the complex scoring methods he reviewed.

Multiple answer items. An alternative method for measuring complex content is to simply create a multiple-choice item with more than one correct answer. Many authors of textbooks and journal articles who prescribe guidelines for writing test items suggest that multiple-choice items should include only a single correct answer (Haladyna & Downing, 1989, Burton et al., 1991). The primary reason for this recommendation is the claim that MA questions

tend to be much more difficult, and as an item's difficulty approaches zero (meaning no one answered it correctly), the item discrimination also tends to be low (Frisbie, 1990). However, Hsu, Moss, and Khampalikit (1984) found MA items to be more discriminating than SA items and provided more information for high-ability examinees.

On the surface, items of this type look very similar to a traditional multiple-choice item (see Figure 2), except that the examinee is instructed to select more than one correct answer (Willson, 1982; Sireci & Zenisky, 2006). The item is scored dichotomously, requiring all correct answers to be selected and no incorrect options to be selected.

Which item properties are used to calculate item reliability? (Select all that apply.) A. Item Discrimination B. Number of Items C. Average Time D. Item Difficultv

Figure 2. Example of a Multiple Select Multiple-Choice item

In most cases found in the literature search, the examinees were simply instructed to "Select all that apply." The examinee is left to decide how many options they should select, which could vary between one and all of the options. Hence, it is easy to understand why these items would be more difficult. Duncan and Milton (1978) present a format where examinees are instructed that there are two correct answers, but their model was not dichotomously scored and never varied beyond two correct answers. Hsu et al. (1984) explored various partial credit scoring methods for MA items, and found that partial credit scoring yielded higher reliability in

the test scores. However, they report that the gains were minimal and not worth the added effort required to apply the scoring model.

Another common problem is that examinees frequently fail to follow the directions to select more than one answer (Pomplun & Omar, 1997). However, it is important to point out that in the studies where it was reported that examinees were confused about whether they were supposed to pick one or multiple options were in paper/pencil administered exams.

In computer administrations, the test sponsor can control the maximum and minimum number of options an examinee will be allowed to select. When these restrictions are enforced, the examinee is not allowed to leave the current item until the correct (or minimum) number of options has been selected. While this feature is not always implemented, it remains an option for audiences that may be more inclined to make this type of error. For example, it may be more beneficial to invoke this option on tests for elementary school children (Pomplun & Omar's 1997) than for certification and licensure exams that target adult learners who have already graduated with at least a bachelor's degree.

In order to reduce the difficulty to a similar level as SA items, David Foster (1999), Robert Frary (1989) and Prometric (2004), a leading computer-based testing company, have recommended that examinees be directed in the stem of each item how many choices they are to select. This is referred to as cueing and is often incorporated in the wording of the text of the stem as well as in a separate comment in parentheses at the end of the stem (See Figure 3).

Which two item properties are used to calculate item reliability? (Choose two.)

A. Item Discrimination
B. Number of Items
C. Average Time
D. Item Difficulty

Figure 3. Example of a Multiple Select Multiple-Choice Item with Cueing

Directing the examinee both in the wording of the stem and at the end of the question, as seen in Figure 3, is referred to as double cueing. Double cueing in the question stem helps to eliminate the error of omission caused by the misunderstanding of examinees as to the number of options they are supposed to pick. Omission of answers can be completely eliminated by invoking the computer delivery option previously described of setting minimum and maximum limits on each item.

Some may argue that instructing examinees exactly how many options are correct lessens the validity of the test scores by providing the examinees with too much of a hint. While this may be true when comparing an MA item with cueing to an MA item without cueing, it not true when comparing MA items to SA items where the examinee knows they are only to select a single answer. It also provides no more of a clue than the alternative of combining all of the correct answers into a single correct response.

Since the purpose of this study is to find an item type that will allow the measurement of complex content that will demonstrate statistical characteristics similar to SA items, then letting the examinee know how many options to select would appear to be an argument for their use because it makes them more similar to their SA counterpart. Additionally, if the statistics of this

item type are good compared to SA items, then the clueing is not an issue, which is precisely what is reported by Foster (1999) and Prometric (2004).

Research Questions

The purpose of this study was to investigate the usability of MA items in high-stakes exams. The study focused on the issues presented in the following four research questions.

Content validity of MA items. How does the content validity of tests composed of MA items designed to measure complex KSAs compare to the content validity of SA items designed to measure the same KSAs?

Acceptance rates of SA and MA items. To what extent do the item analysis statistics for MA items comply with the generally-accepted quality standards for deciding which item to retain on an exam?

a. The preferred standard is for items to have point-biserial correlations greater than .30, item difficulties between .25 and .80, and item reliabilities greater than .125.
b. Items that fail to meet the preferred standard but do meet the minimum standard are considered marginal. The minimum standard is for items to have point-biserial correlations greater than .20, item difficulties between .15 and .90, and item reliabilities greater than .09.

Average statistical characteristics of SA and MA items. How do the average statistical characteristics of MA items compare to SA items in terms of the following?

a. Item difficulty
b. Item discrimination
c. Item reliability
d. Item information

Partial credit versus dichotomous scoring of MA items. How does the use of partial credit scoring compare to dichotomous scoring where the possible point total for any given item is equal to 1.0 point?

a. the distribution of the resulting scores

b. the estimated reliability of the scores

c. the test information functions

Chapter 2: Literature Review

Traditional multiple-choice items have one correct answer and three or four incorrect answers. Because of the need to measure complex content that often has more than a single answer, various item types have been introduced such as Type K items, multiple true false (MTF) items (also known as Type X items), and multiple select (MA) items.

It is important to distinguish what I have classified as an MA item from items that instruct the examinee to select the best answer from among a list of options that include more than one correct answer, but one of the correct answers is clearly better than all others. In one sense these items have more than one correct answer, but they require the examinee to select only the single best answer. While some test development experts advocate the use of best option format (see Haladyna & Downding, 1989), I have not included this item type with the MA items.

Single Correct Answer Recommendation

Some scholars, such as Shrock and Coscarelli (2007), recommend avoiding MA items. Burton, Sudweeks, Merrill, and Wood (1991) also counsel against the use of multiple correct answers in items citing research that indicates these items are lower in reliability, higher in difficulty, and equal in validity when compared to similar items that have a single correct answer, or best answer format (Frisbie, 1990).

Haladyna and Downing (1989) reviewed 46 authoritative references in the field of educational measurement and found that 35 of the 46 addressed how many correct answers should be included in an item. Thirty-two of those thirty-five sources recommended using only one correct answer. These references ranged from publications as early as 1936 (Hawkes, Lindquist, & Mann, 1936) and as late as 1987 (Kubiszyn & Borich, 1987). In this review,

Haladyna and Downing (1989) classified each rule as either a value-laden or an empirically testable rule (p. 46). A value-laden rule was operationally defined as "a rule that is generally regarded by measurement experts and practitioners as acceptable advice without further question or discussion" (p. 45). They go on to say that because consensus exists for the validity of the rule, that no empirical research should be anticipated or planned to validate the rule. Since most of the rules thus classified are general guidelines that would either be difficult to test or so obvious to the experts as sound advice, such advice seems appropriate. For example, some of these value rules include: (a) "Use grammatical consistency," (b) "Avoid verbatim phrasing," (c) "Avoid opinions," and (d) "Use plausible distractors." It is understandable why these rules would be considered good advice and unnecessary to test.

However, it is hard to understand why the rule to "Include one correct answer" would have been categorized as such. What it does indicate is that the advice of the 32 authors of this item writing guideline did not base their advice on empirical data, but rather on what they believed at the time to be a good idea. However, it is not clear why Haladyna and Downing considered this rule to be untestable. In another review of literature, Haladyna, Downing and Rodriguez (2002) reported that because the rule for selecting one right answer was unanimously endorsed by all authors in their review, it was accepted as a "common core" item writing guideline and no further consideration was given to it.

In a more recent publication, Haladyna (2004) writes very positively about MA items by referring to how well the performance of these items compare to other multiple-choice formats. His prediction is that this format will become more widely used in both classroom and formal, standardized testing programs as more research is completed on this type of item.

Type K Items

The most criticized item type that addresses complex content has been the Type K items. While this item type only requires a single response from the examinee, it does require the examinee to know more than one component of the complex content in order to answer the question. This item type generally presents three or more answers to the examinee, but the options the examinee is allowed to choose from are different pre-determined combinations of those answers.

This item type is used in many certification and medical licensure exams. In the comparative review by Haladyna and Downing (1989), eight of nine studies they reviewed reported that Type K items were harder and exhibited lower discrimination than single-answer (SA) items. That Type K items are harder is no surprise, since there is more information being tested in the items than in SA items. Since no credit is given for partial knowledge, an examinee with some knowledge is scored the same as an examinee with no knowledge. However, in content areas where this type of scoring is appropriate, these items will likely have higher discrimination and reliability. For example, if there is no value in knowing part of the information, why give credit for the part that an examinee knows? This would explain why this item type is more commonly used in medical licensure exams and not in educational settings. In healthcare, a patient is not likely to care if a physician knows only one of the two pre-operative safety procedures.

Albanese (1982) claims that Type K items offer clues to test wise examines who can identify answer options they know to be false, thus helping them narrow the choices down to only a few. For example, consider the item shown in Figure 1. If the only information an examinee knows about that item is that item content is not a correct answer, the examinee will

have a 50% change of guessing because item content is in both options A and C. Since options B and D both include "Item Variance" as an answer, we have given the examinee half of the answer, so we are really only testing whether "Item Discrimination" or "Item Kurtosis" is a correctly answer. With four options and two correct answers, the chance of guessing should only be 17%, if the examinee were to identify their own combinations. As stated by Albanese (1982), "One trivial false option turns a Type K item into essentially a true-false item" (p. 224).

Table 1 shows what the guessing probabilities would be for various numbers of options and correct answers if all combinations of answers were presented to an examinee. Type K items limit the number of combinations by pre-determining which answers will be presented together, thus making it easier to eliminate some options with only a partial knowledge.

Multiple True-False Items

An alternative to Type K items is the Multiple True-False (MTF) item, sometimes referred to as a Type X item. Albanese (1982) compared MTF items, scored dichotomously, to Type K items and found that the Type K items were substantially easier and less reliable than MTF items. Harasym, Norris, and Lorscheider (1980) also found Type K items to be easier and suggested that this could be related to the clueing effect of the limited combinations as previously discussed. Based on the evidence of clueing and difficulty, several researchers, including Albanese (1982), Harasym et al. (1980), and Carson (1980), recommend the MTF item over the Type K item.

Table 1

Random Guessing Probabilities for Multiple-Choice Items

Number of Options Presented	Number of Correct Answers	Number of Possible Combinations	Chance of Guessing
3	1	3	33%
3	2	3	33%
4	1	4	25%
4	2	6	17%
4	3	4	25%
5	2	10	10%
5	3	10	10%
5	4	5	20%
6	2	15	7%
6	3	20	5%
6	4	15	7%
6	5	6	17%

Notwithstanding the findings that support MTF items as superior to Type K items, the concern about the increased difficulty of Type K items is exacerbated with MTF items. As the previous literature review shows, while MTF items are more reliable than Type K items, they are significantly more difficult. As a result, the usefulness of these items in a practical setting is also diminished.

Albanese and Sabers (1978) evaluated four methods for scoring MTF items. The levels varied from scoring them dichotomously, as previously reported, to scoring each option independently as a true-false item. When each option was scored independently, they found that in most cases, not only was reliability improved, but the items were much easier because partial credit was given for partial knowledge. Also, the difficulty of constructing Type K items is greatly simplified by using MTF items and the cognitive demand on examinees is lessened, enabling more items to be presented in a set amount of time (Huntley & Plake, 1984).

Another difficulty in using MTF items in computer-administered exams is the limitation of the test drivers (the computer program that administers the exams) of the major test delivery vendors, where most do not have the ability to present and score a MTF item. Because the MA format is so much easier to administer in the computerized testing environments, and the fact that they are closely linked to MTF questions, this is the format most often utilized by certification and professional licensure programs that deliver through the major computer delivery companies. The only difference between the MA item and the MTF item is that the later requires the examinee to specifically mark each option as correct (true) or incorrect (or false), whereas the former only requires the examinee to mark the options they believe are correct and omit making any marks on incorrect options. Cronbach (1941) evaluated the performance difference between MA items and MTF items and indicated that his evidence supported the use of MA items over

MTF if unintentional omissions were not expected. One of the reasons for his recommendation is what he calls the hypothesis of acquiescence, where students have a tendency to mark MTF options as true rather than false when guessing or if they are not sure.

Multiple Answer Items

The format of the MA item type most closely resembles that of the traditional SA item, but it allows for more than one correct answer to be selected, thereby closely approximating the advantages of MTF items. The examinee is either told how many correct choices to select or is instructed to select all that apply. The most common format is to instruct the examinee to select all options that apply, without indicating whether there is only one or if there are multiple correct answers. Common criticisms of this item type are that they (a) tend to be more difficult, (b) have low discriminating power, and (c) take more time to answer.

Willson (1982) found that the reliability of a multiple-choice test can be improved by allowing several or all of the options to be correct. He treated each option in an item as a subtest with N parallel forms, with N being the total number of options in the item. In essence, he scored the items the same as he would have scored a MTF item without requiring the examinee to explicitly mark options as false. This is also analogous to the more traditional MA items with partial credit being applied. LaDuca, Downing, and Henzel (1995) indicated that in some contexts, MA items are preferable over SA items for licensure and certification exams. For example, "when equally attractive treatment options may exist for selected illnesses, or several appropriate diagnostic studies should be pursued" (p. 121).

Hsu et al. (1984) compared the performance of MA and SA items. Their results indicated that the reliability of MA items was equal to SA items but that MA items were consistently more discriminating than SA items. While the MA items were generally more difficult than SA items,

since higher ability examines were able to answer these items correctly, these items would be preferred on exams that are intended to identify medium to high-ability individuals. Also, in an earlier study by Dressel and Schmid (1953) MA items were found to have higher reliability than SA items.

Summary

While most early (1930's through mid-1980's) researchers recommended only including a single correct answer in multiple-choice items, this recommendation was not based on reported empirical research (Downing & Haladyna, 2006). The combined recommendation from these early publications was classified as a value rule where no empirical research was conducted to validate the rule. However, subsequent research has identified several item types that allow for more than a single answer that result in acceptable reliability and discrimination. The most controversial of these item types is the Type K item, but as previously mentioned, MTF items have been overwhelmingly preferred over the Type K items. An alternative to MTF items is the MA items, where partial credit is given.

Most research studies involving MA items do not include MA items that prompt the examinee for how many options they are to select. These items have been used in the certification and professional licensure for over 10 years (Foster, 1999), but very little has been published about their performance compared to other item types. The reason this alternative for MA items is appealing is that it allows for the current computerized delivery technology to be utilized. It is well-suited for computer administration because computerized delivery engines can prevent an examinee from selecting too many or too few options, thus reducing errors made by the examinee who either did not understand the instructions or did not realize that they were expected to select more than one answer. Additionally, by cueing examinees for how many

answers they are expected to select, it is hypothesized that the difficulty of these items will be reduced, bringing them closer to the level of the traditional SA items. As long as these items are properly applied, it is expected that this will also have a positive impact on both the reliability and discrimination of the items as well. A proper application would be a situation where there is no advantage in having partial knowledge over no knowledge of the domain covered in an item.

Chapter 3: Method

The Selection of Exams Used in This Study

Three operational certification exams were used in this study. Two are certification exams for Information Technology (IT) companies that measure IT professionals' proficiency in working with and utilizing their respective software and hardware products. The third is a certification exam for a professional association that measures a much broader set of skills that apply to their profession as a whole. The following criteria were used to select the three sample tests analyzed in this study.

1. Diversity of the content measured. While two of the exams were IT certification exams, they covered completely different content areas. The third exam was a job certification hosted by a professional association. This variety of content was intended to provide a broader comparison for SA and MA items across different content areas.

2. Diversity in the knowledge, skills, and abilities (KSAs). The KSAs being assessed ranged from highly detailed and explicit to very broad and less well-defined (or documented). The two IT certification exams are used for making product certification decisions, so the focus of these exams is quite narrow and well-defined. However, many job certification tests are not based on universally-accepted criteria, but are based on best-practice, specific documents or specific training materials. But even with these materials as a basis, the nature of the KSAs for some job certifications, including one of the tests used in this study, makes it difficult to measure with discrete, selected response questions. The result is often lower-correlating items and lower item reliability. So, the purpose of including one of these exams in the study was to determine if the relative performance MA and SA items are affected by this type of content.

3. Minimum number of response records. A minimum of 300 examinee test records was required for each exam in this study. This requirement was established to provide stability in the Classical Test Theory statistics and to provide sufficient data to perform a one-parameter IRT analysis for all of the exams.

4. Adequate number of SA and MA items. Each exam was comprised of a minimum of 50 multiple-choice items, with at least 30% of the items being comprised of MA items. The exams needed to have sufficient items of both item types to perform a meaningful comparison between MA and SA items.

Test records for examinees who did not answer all questions within an exam and for examinees who spent less than five seconds responding to over 5% of the items were deleted from this study. The reason for omitting examinees who did not answer all questions was because it was not possible to know if the examinees did not answer an item because they did not know the answer, because they ran out of time, or because they may have simply skipped it and forgot to come back to it. Even though the software can effectively deal with the missing data, Classical Test Theory relies on the total scores as the examinees' ability estimate, and if an examinee did not answer an item for which they actually knew the answer, the ability estimates based on their total score would be incorrect.

The reason for omitting people who spent less than five seconds on over 5% of the exam was the concern that they were simply marking answers without even reading the items. Some item could very well be very short and simple to answer, which is why I allowed for at least 5% of the items to be quickly answered, but it was felt that if the number of items exceeded 5% for a given examinee, it was likely due to marking answers without reading the item. Table 2 shows the number and percentage of records that were omitted for each exam.

Table 2

Examinee Test Records Omitted From This Study

	Number of	Omitted	
Exam	Respondents	Number	Percent
1	661	23	3.5%
2	3,322	281	8.5%
3	490	25	5.1%

Design

Content validity of MA items. To determine the content validity of the MA items, an external validation study was conducted, where three subject matter experts (SMEs) for each exam reviewed the content of the MA items for their assigned exam. The raters were presented with four rating tasks for each MA item in their respective exams. Each rating task was provided the SMEs the opportunity to rate how appropriate the use of each dichotomously-scored MA items was for the testing objective being measured and for the ability level of the target audience being assessed.

The experts were able to view each MA item and its assigned testing objective together to perform the four rating tasks for each item. The concept of validity is that an item measures what it purports to measure. Therefore, to establish the validity of the MA items, the SMEs were asked to rate how necessary it was to require more than one answer in order to adequately measure all of the requirements of the objective and whether or not scoring the items dichotomously was appropriate.

Each SME was instructed on the use of MA items and SA items and the implication of each in certification testing. For example, the SMEs were told that the examinees will not receive any credit for partial knowledge and were asked to rate how appropriate that scoring method is for the content of each question. The reviewers were told that the initial item writers and reviewers were instructed that if they created an item with more than one correct answer, that the content should be such that an examinee would need to know ALL answers in order to demonstrate minimal competency for that content area. The reviewers were asked to rate how well each of the MA items met that criteria.

To protect the security of these operational exams, only the MA items were shown to the SMEs. Since the test sponsors (owners of the test content) were very concerned about sending copies of items on their current active exams out for review, a software application was used that allowed the SMEs to log into a secure website to perform their reviews. To access an exam, they were required to provide an event URL that contained a 32-character hexadecimal number and an event access code. After the successful entry of these two pieces of information, they were then required to log into the tool using a unique username and password that we set individually for each SME. The SMEs were asked to complete the following four item rating tasks for each MA item in each of their respective tests.

1. The examinee will get one point by selecting ALL correct answers and NO points for selecting a portion of the correct answers.

   How appropriate is this scoring method for the content being measured by this item?

   A. Not at all appropriate

   B. Somewhat inappropriate

   C. It doesn't matter

D. Somewhat appropriate

E. Very appropriate

2. How much value is there in the examinee knowing part, but not all, of the correct answers with respect to determining minimal competency for this exam?

   A. Very high value in partial knowledge

   B. High value in partial knowledge

   C. Some value in partial knowledge

   D. Low value in partial knowledge

   E. Very low value in partial knowledge

3. How much better could the knowledge, skills, abilities, or judgments covered by this question be measured with a single-answer multiple-choice item?

   A. Much better

   B. Somewhat better

   C. The same

   D. Somewhat worse

   E. Much worse

4. How important is it for the examinee to know all of the correct answers in this question in order to demonstrate minimal competency in the domain covered by this exam?

   A. Not important

   B. Somewhat important

   C. Important

   D. Very important

Figure 4 shows a screen shot of the tool used by the SMEs to view the items and to complete the rating tasks for each item. The SMEs for each exam completed the rating tasks by selecting options A through E for each of the four tasks about each MA item. In order to summarize their responses, their ratings were converted to numeric values by the following assignments: for questions one through three, A = 0, B = 1, C = 2, D = 3, and E = 4 and for the fourth question, A = 0, B = 1, C = 3, and D = 4. The reason for coding the fourth question differently was because that question did not have a neutral option and the goal was to have all questions on the same scale, with values closer to zero indicating an inappropriate application of dichotomously-scored MA items and values closer to one indicating an appropriate application of dichotomously-scored MA items.

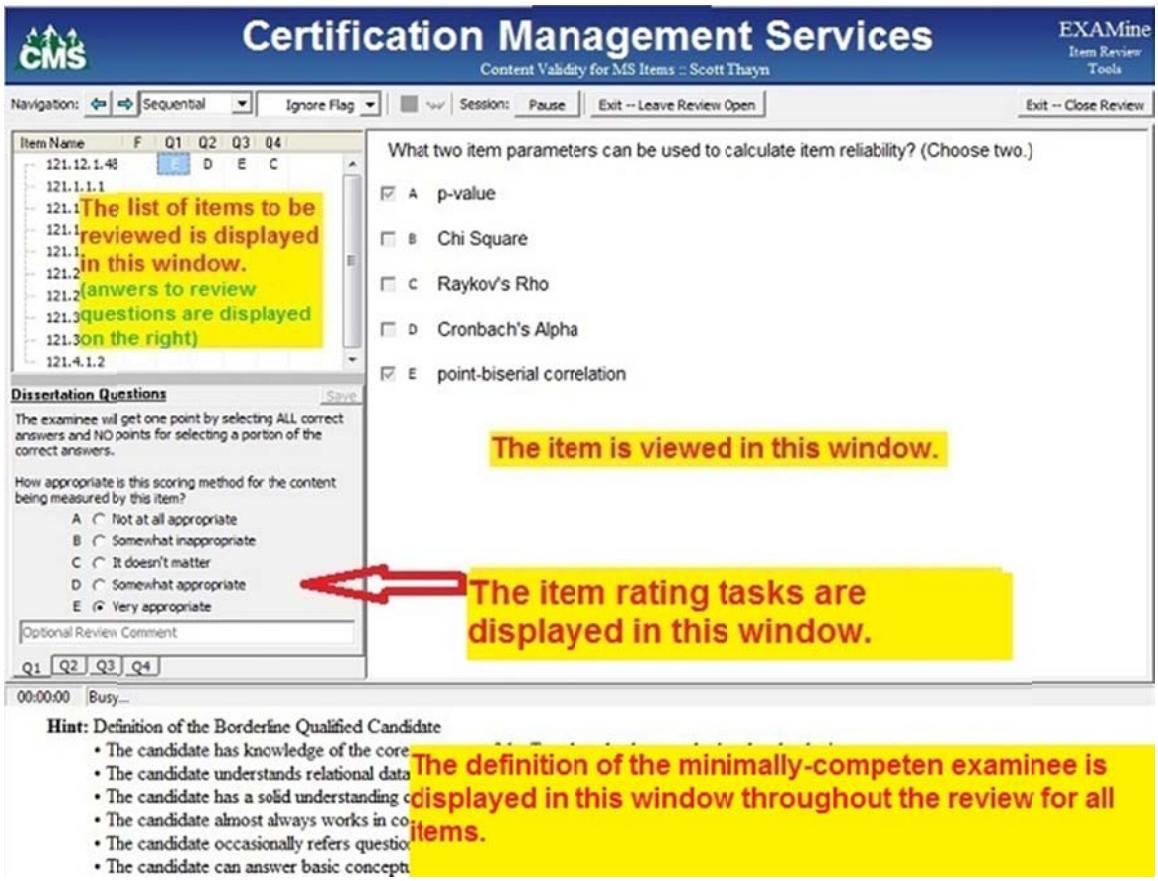

Figure 4. Screen shot of review tool.

Acceptance rates of SA and MA items. A Classical Test Theory analysis was conducted on all three exams. An item discrimination index, an item difficulty index, and an item reliability statistic were calculated for each item within each exam. The items were categorized as SA or MA items and the percentage of items within each category that met the Minimal and Preferred standards were calculated. The standards used to judge the items are shown in Table 3. The percentage of SA and MA items within each exam that met each of the individual criteria listed in Table 3 was calculated as well as the total combined percentages of all SA and MA items across all exams.

Table 3

Statistical Standards

Statistic	Minimal	Preferred
Item Discrimination	.20	30.0%
Item Difficulty	.15 ≥ P ≥ .90	.30 ≥ P ≥ .75
Item Reliability	.10	15.0%

The item discrimination statistic is a correlation of the dichotomous item scores to a total score for all examinees for a given exam. The total scores are used as an estimation of the overall ability level of the examinees for the domain covered by the exam. This correlation indicates whether or not high-ability examinees have a higher tendency to answer a question correct than low-ability examinees. The range for this statistic is -1.0 to +1.0, with +1.0 indicating that the half of the examinees that scored the highest on the exam all got a given item correct and the half that scored the lowest on the exam all missed that item.

The item difficulty index is the proportion of examinees who answered a given question correct. The range for this statistic is zero to 1.0, with numbers close to zero indicating hard items and numbers close to 1.0 indicating easy items.

The item reliability statistic mathematically combines both the item difficulty and the item discrimination for a given item. Item reliability is calculated by multiplying the item discrimination index by the item standard deviation and is given by the following formula:

$$R_i = r_{it}\sqrt{pq}$$

Where:

$R_i =$ Item Reliability

$r_{it} =$ Item Discrimination Correlation

$p =$ Item Difficulty

$q = 1 - p$

This statistic is an indication of the overall quality of items. For example, it can be used to determine if an item with a strong discrimination index and a minimally-acceptable item difficulty index is better than an item with a moderate discrimination index and a moderate item difficulty index. Additionally, it can be used to determine if an item meets minimal item discrimination and minimal item difficulty standards, then the item may not be helpful in determining pass/fail status on an exam. Therefore, as additional acceptance criteria, minimal and preferred standards are often used for this statistic as well.

The purpose of this procedure was to evaluate how well each of the item types met the general acceptance criteria for inclusion on a high stakes exam compared to the other. This will be evaluated by calculating the percentage of SA and MA items to meet the criteria.

Average statistical characteristics of SA and MA items. To address this research question, the item correlation indices were calculated two ways. First, the examinees' responses to each item were correlated with their total score on the whole test which included their

combined score on both the SA and MA items. Second, the examinees' responses to each SA item was correlated with their total score on only the SA portion of the test, and their responses to each MA item was correlated with their total score on only the MA portion of the test. This was done to separate the influence of the performance of one item type when calculating the statistics of the other item type.

Since the correlation index for an item is used to calculate the item reliability index, two separate item reliability indices were calculated for each item as well. The means and standard deviations were calculated for each statistic by item type within each exam. This design allowed for not only comparing the average SA item performance to MA item performance, but it also provided information about how each item type performed independent of the other item type's influence.

Finally, the Winsteps® software was used to estimate the Rasch difficulty parameter for each item. Using these difficulty estimates, the item information functions were generated using the following formula.

$$I_i(\theta) = \frac{2.89}{e^{1.7(\theta - b_i)}(1 + e^{-1.7(\theta - b_i)})^2}$$

Where b_i is the item difficulty, and θ is the ability level.

A spreadsheet was created for each exam to calculate the item information for each item at 25 different ability levels, ranging from a theta of -3.0 to a theta of +3.0 at .25 increments. The item information was then averaged for all SA and MA items within each exam for each ability level. The average item information functions for SA and MA items were generated and graphed by exam.

The Rasch model assumes that all items are equally discriminating and that they only vary in difficulty. However, based on the previous Classical analysis, the items in all three

exams demonstrated widely-varying discrimination. Therefore, to determine if accounting for the differences in item discrimination would result in a visible difference between the average SA and MA item information functions, the discrimination estimates generated by Winsteps were used to re-calculate the average item information functions for SA and MA item using the following two-parameter IRT item information formula.

$$I_i(\theta) = \frac{2.89a_i^2}{e^{1.7a_i(\theta - b_i)}(1 + e^{-1.7a_i(\theta - b_i)})^2}$$

Where a_i is the item discrimination, b_i is the item difficulty, and θ is the ability level.

Partial credit versus dichotomous scoring of MA items. To answer the last research question, two Rasch analyses were conducted on each exam. One Rasch analysis was performed on the test results where the items were scored dichotomously and the second was performed on the test results where the items were scored using partial credit scoring.

The first task was to score the exams such that the results of all analyses would be on the same scale and in such a way as to keep the relative weighting of each item the same. To meet the strict requirements of the exam blueprints, all items must carry the same scoring opportunity (one point per item), regardless of whether the item type is SA or MA. Also, it was desirable to use whole number scoring to facilitate the use of Winsteps to perform the analysis. To further complicate the scoring requirements, the number of correct options in the MA items varied from two correct to four correct in all of the exams. To accommodate all of these requirements, it was decided to score all items on a 12-point scale.

For the dichotomously scored analysis, examinees were given 12 points for each item in which they answered all options correct and no points if they did not correctly identify all correct answers. For the partial credit analysis, examines were scored based on the total number of correct options in an item. If two options were correct, they received six points for each correct

option they selected. If three options were correct, they received four points for each correct option they selected. And finally, if four options were correct, they received three points for each correct option they selected.

The fit statistics for each analysis were reviewed and compared between the dichotomous scoring and the partial credit scoring to assess the goodness of fit of the data to the models. Once it was established that the model fit was acceptable for each model, reliability coefficients were compared between the two scoring models.

Next, the item information was calculated for all items, for both scoring models, for all three exams. The same process described in research question 3 was used to generating the item information across all ability levels. To provide additional insight, the average item information was grouped by item type to determine if the scoring model had more impact on one item type than the other.

Chapter 4: Results

Content Validity of MA Items

The mean scores and standard deviations for each rater on each survey question were calculated and are reported in Table 4. The total mean scores and standard deviations for the item rating tasks were then calculated by combining all raters and all questions for each item rating task within each exam. The task means and standard deviations are displayed in Table 4 under the sub-heading labeled Total for each item rating task for each exam. The grand mean and standard deviation combines the responses of all raters to all survey questions for all items within each exam.

The grand mean is the overall measure for how appropriate the MA items were applied to an exam. Values that are closer to 1.0 indicate a higher appropriateness for the use of dichotomously-scored MA items for their given objectives. The grand standard deviation indicates the variability of all ratings within each exam.

The grand means of .86 for Exam 1 and .74 for Exam 3 indicate an overall strong approval by the SMEs of the use of MA items on those exams. The lowest variability for all combined ratings was observed in Exam 1 (.20), which indicates the most consistent ratings between raters and their opinions about the appropriateness of the items.

Exam 2 only scored a grand mean of .56 with a higher grand standard deviation of .28. The low mean indicates only a moderate approval for the use of these dichotomously-scored MA items on that exam. However, based on many of the comments made by the SMEs, one of the reasons for the lower ratings resulted from a concern about the overall quality (and sometimes even the technical accuracy) of the items, rather than the number of correct responses. This

Table 4

Summary of Item Rating Tasks

		Item Rating Task 1				Item Rating Task 2				Item Rating Task 3				Item Rating Task 4				Grand	
Exam	Statistic	R1[a]	R2	R3	Total	R1	R2	R3	Total	R1	R2	R3	Total	R1	R2	R3	Total	Mean	SD
1	Mean	.98	1.00	.84	.94	.73	.98	.56	.75	.93	.98	.76	.89	.81	.97	.78	.86	.86	.20
	SD	.07	.00	.20	.14	.19	.10	.20	.24	.14	.06	.15	.16	.20	.09	.22	.20	.07	
2	Mean	.99	.59	.39	.65	.66	.56	.28	.50	.74	.51	.40	.55	.68	.48	.43	.53	.56	.28
	SD	.05	.24	.24	.32	.18	.19	.22	.26	.13	.22	.19	.23	.20	.31	.24	.28	.06	
3	Mean	.68	.95	.86	.83	.53	.78	.60	.63	.62	.80	1.00	.81	.53	.85	.64	.67	.74	.28
	SD	.26	.16	.18	.23	.31	.23	.24	.28	.25	.24	.00	.25	.32	.24	.28	.31	.08	

[a]Note. R1, R2, and R3 designate Raters 1, 2 and 3 respectively.

exam is over five years and several product revisions old now, even though it is still an operational exam.

Exam 3 resulted in a grand mean of .74, which indicates a high overall approval of the MA items. However, there was a high variance (.28) in the all of the combined ratings for this exam.

The standard deviations listed underneath the grand means are the variances in the means for each item rating task. Each item rating task was intended to obtain the same information from the SMEs, so the variance in the mean responses for each item rating task was expected to be low. As shown in Table 4, these variances were relatively low, with standard deviations ranging from .06 to .08 for all exams. This indicates that the SMEs rated each of the item rating tasks within a given item very similarly. Therefore, the high total (grand) variances experienced in Exam 2 and Exam 3 were primarily due to the variance between raters and items.

Acceptance Rates of SA and MA Items

Table 5 shows the percent of SA items that met both the minimal acceptable item difficulty criteria ($.15 \leq$ difficulty index $\leq .90$) and the preferred item difficulty criteria ($.30 \leq$ difficulty index $\leq .75$). In addition to listing each exam independently, the total percentage was calculated across all exams.

Overall, a higher percentage of MA items met both the minimal and the preferred standards. On an exam-by-exam basis, the SA items resulted in a higher acceptance rate for only one of the three exams at the minimum standard. In all other comparisons, MA items had a higher acceptance rate.

Table 5

Percentage of SA and MA Items That Satisfy Minimal and Preferred Difficulty Standards

	Minimal Standard (.15 ≤ Difficulty ≤ .90)				Preferred Standard (.30 ≤ Difficulty ≤ .75)			
	SA		MA		SA		MA	
Exam	Ratio	Percentage	Ratio	Percentage	Ratio	Percentage	Ratio	Percentage
1	47/49	96%	61/62	98%	35/49	71%	47/62	76%
2	25/44	57%	32/44	73%	4/44	9%	7/44	16%
3	113/129	88%	50/59	85%	59/129	46%	37/59	63%
Total	185/222	83%	143/165	87%	98/222	41%	91/165	55%

Table 6 displays the same type of comparison for the item discrimination indices, where the minimally acceptable value is .20 and the preferred minimum is .30. The ratios and percentages are reported separately for each exam and for all three exams combined.

Overall a higher percentage of MA items met both the minimal standard and the preferred standard. On an exam-by-exam basis, the SA items had a slightly higher percentage of items meet the minimal standard of .20, but a higher percentage of the MA items meet the preferred standard of .30.

Table 7 is a summary of the percentage of both SA and MA items that met the minimal standard of .10 and the preferred standard of .15 for item reliability by exam and combined across all exams. The total acceptance rate for MA items was 20% higher for the minimum standard and 24% higher for the preferred standard.

Table 6

Comparison of the Item Discrimination Acceptance Percentages

	Minimal Standard (Point Biserial ≥ .20)				Preferred Standard (Point Biserial ≥ .30)			
	SA		MA		SA		MA	
Exam	Ratio	Percentage	Ratio	Percentage	Ratio	Percentage	Ratio	Percentage
1	48/49	98%	61/62	98%	42/49	86%	52/62	84%
2	44/44	100%	42/44	95%	38/44	86%	39/44	89%
3	58/129	45%	23/59	39%	3/129	2%	4/59	7%
Total	150/222	68%	126/165	76%	83/222	37%	95/165	58%

Table 7

Comparison of the Item Reliability Acceptance Percentages

	Minimal Standard (Item Reliability ≥ .10)				Preferred Standard (Item Reliability ≥ .15)			
	SA		MA		SA		MA	
Exam	Ratio	Percentage	Ratio	Percentage	Ratio	Percentage	Ratio	Percentage
1	46/49	94%	61/62	97%	36/49	73%	52/62	69%
2	32/44	73%	33/44	75%	17/44	39%	23/44	52%
3	33/129	26%	21/59	36%	1/129	1%	4/59	7%
Total	111/222	50%	115/165	70%	54/222	24%	79/165	48%

The acceptance rate for MA items within each exam was higher for all exams at the minimum standard and for two of the three exams at the preferred standard. Only Exam 1 demonstrated a higher acceptance rate (4%) for SA items than MA items.

Average Statistical Characteristics of SA and MA Items

Table 8 provides a summary of the average Classical Test Theory statistics for each exam and a summary of all items from all exams combined. The discrimination indices were calculated two different ways. Therefore, under the column heading Statistic, there are two separate values for the Discrimination indices, which are described in the Description column as Item-to-total correlation and Item-to-subtest correlation.

The item-to-total correlation coefficients were calculated by correlating the examinees' responses to each item to the examinees' total scores within each exam. These total scores were based on examinees' combined scores for both SA and MA items within each exam.

The item-to-subtest correlation coefficients were calculated by creating two subtests for each exam. One subtest was comprised of all of the SA items within an exam and the other subtest was comprised all of the MA items within an exam. Total subtest scores were generated for all examinees for the subtests within each exam. The examinees' responses to each item were then correlated to the examinees' total subtest scores that included that item. Therefore, the discrimination indices for SA items were based only on the examinees' total score for just the SA items and similarly for the MA items.

Since item reliability is partially calculated from the discrimination index, there were two average item reliability indices calculated for each item as well. The average reliability index based on the discrimination indices using total scores is described in the table as Item-to-total

Table 8

Average Item Analysis Statistics by Exam and Item Type

Exam	Statistic	Description	SA Items Mean	SA Items SD	MA Items Mean	MA Items SD	Percent Increase (Decrease)
1	Difficulty index	Percent correct	.660	.143	.547	.181	-21%
	Discrimination indices	Item-to-total correlation	.381	.076	.375	.075	-2%
		Item-to-subtest correlation	.412	.071	.391	.071	-5%
	Item reliability indices	Item-to-total reliability	.170	.039	.174	.041	2%
		Item-to-subtest reliability	.184	.039	.181	.041	-2%
2	Difficulty index	Percent correct	.869	.071	.839	.092	-4%
	Discrimination indices	Item-to-total correlation	.409	.100	.437	.115	6%
		Item-to-subtest correlation	.425	.105	.453	.113	6%
	Item reliability indices	Item-to-total reliability	.138	.063	.160	.073	14%
		Item-to-subtest reliability	.143	.066	.165	.073	13%
3	Difficulty index	Percent correct	.709	.177	.628	.210	-13%
	Discrimination indices	Item-to-total correlation	.176	.082	.174	.094	-1%
		Item-to-subtest correlation	.183	.080	.203	.087	10%
	Item reliability indices	Item-to-total reliability	.074	.039	.076	.048	4%
		Item-to-subtest reliability	.077	.038	.090	.048	15%
All Exams	Difficulty index	Percent correct	.730	.170	.653	.210	-12%
	Discrimination indices	Item-to-total correlation	.268	.137	.320	.146	16%
		Item-to-subtest correlation	.282	.143	.341	.138	17%
	Item reliability indices	Item-to-total reliability	.108	.061	.135	.070	20%
		Item-to-subtest reliability	.114	.064	.144	.068	21%

reliability, and the average reliability index based on the discrimination indices using subtest scores is described in the table as Item-to-subtest reliability.

Classical Test Theory comparisons of SA and MA items. The last column reported in Table 8 shows the percent increase or decrease in the means of the MA items relative to the SA means. The values were calculated by dividing the mean values for SA items by the mean values for MA items and subtracting that value from one.

$$\text{Percent Increase (Decrease)} = 100 * \left(1 - \frac{\text{MA Mean}}{\text{SA Mean}}\right)$$

A positive values in the last column of Table 8 indicates how much higher (expressed as a percentage) an MA item mean was than the SA item mean. Conversely, a negative value indicates how much lower an MA item mean was than the SA item mean.

In Table 8, the average item difficulties for all three exams are lower for MA items than for SA items, meaning that MA items are harder than SA items. As a group, MA items were 12% harder than the SA items, but it is not possible to know if this increase in difficulty is the result of the item type itself or the nature of the test content being measured. However, the difficulty values are well within the preferred range of .30 to .75, so the increased difficulty is not enough to warrant avoiding the use of MA items.

The average item-to-total correlation for MA items was 2% lower than the SA items on the first exam, 6% higher on the second exam and 1% lower on the third exam. With only three exams in this study, these results indicate no difference in the discrimination between SA and MA items.

Rasch analysis of SA and MA items. The mean item information functions for SA and MA items were first calculated using only the Rasch item difficulty estimates. The resulting plots are shown in Figures 5, 6 and 7.

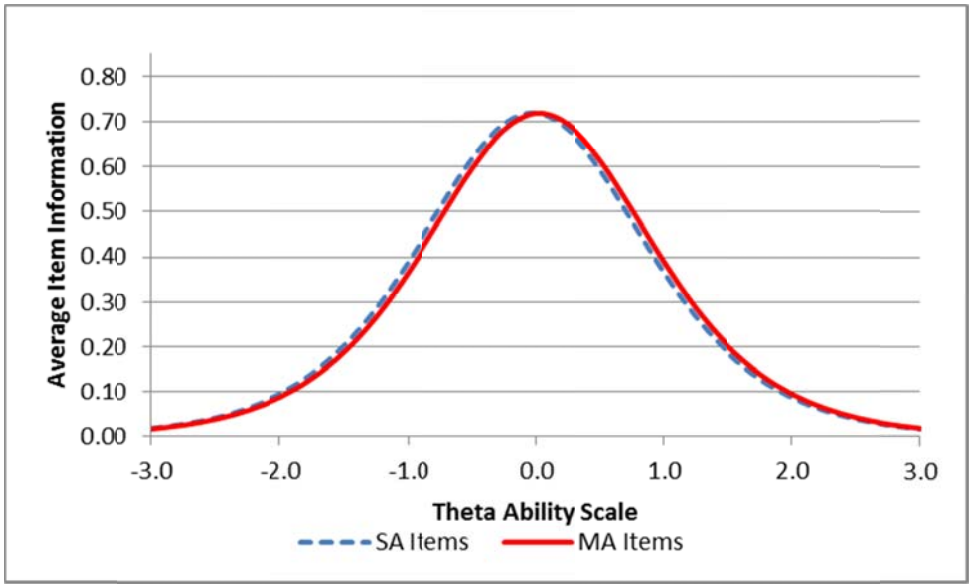

Figure 5. Average item information function for Exam 1 using only Rasch

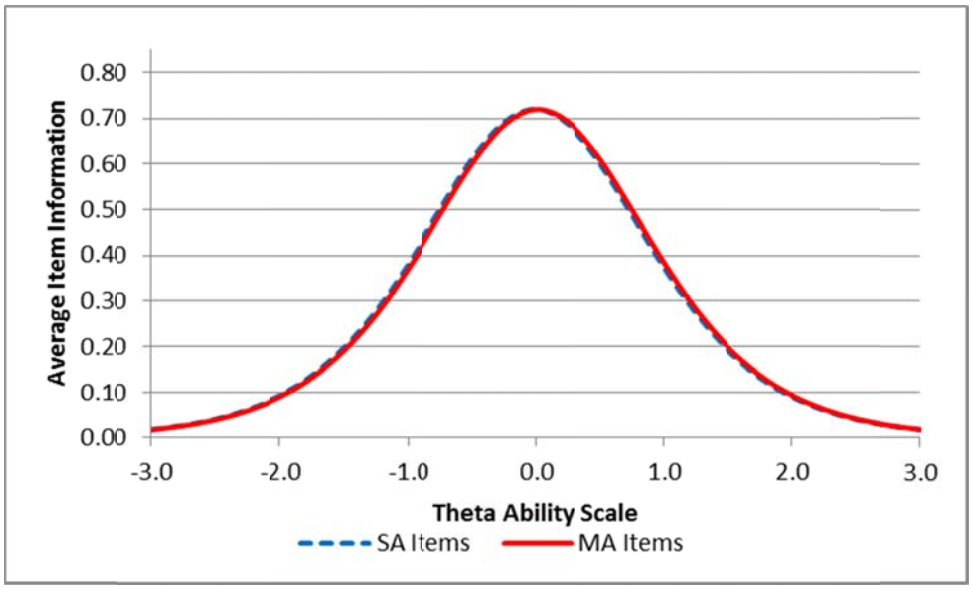

Figure 6. Average item information function for Exam 2 using only Rasch

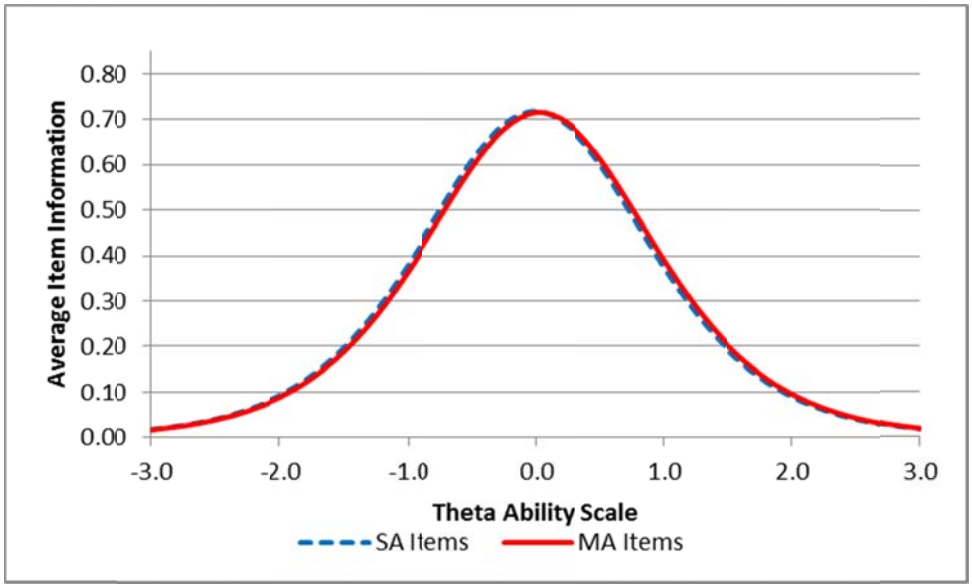

Figure 7. Average item information function for Exam 3 using only Rasch

Based on the near overlapping curves, SA and MA items resulted in virtually the same average item information across all ability levels. The mean item information functions for SA and MA items using two-parameter formula are shown in Figure 8, Figure 9 and Figure 10.

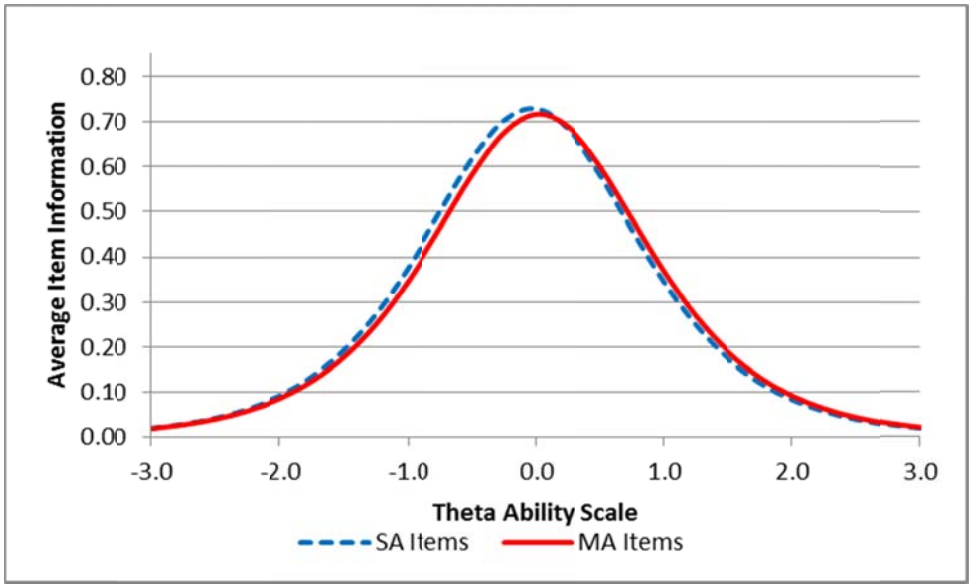

Figure 8. Average item information function for Exam 1 using two parameters

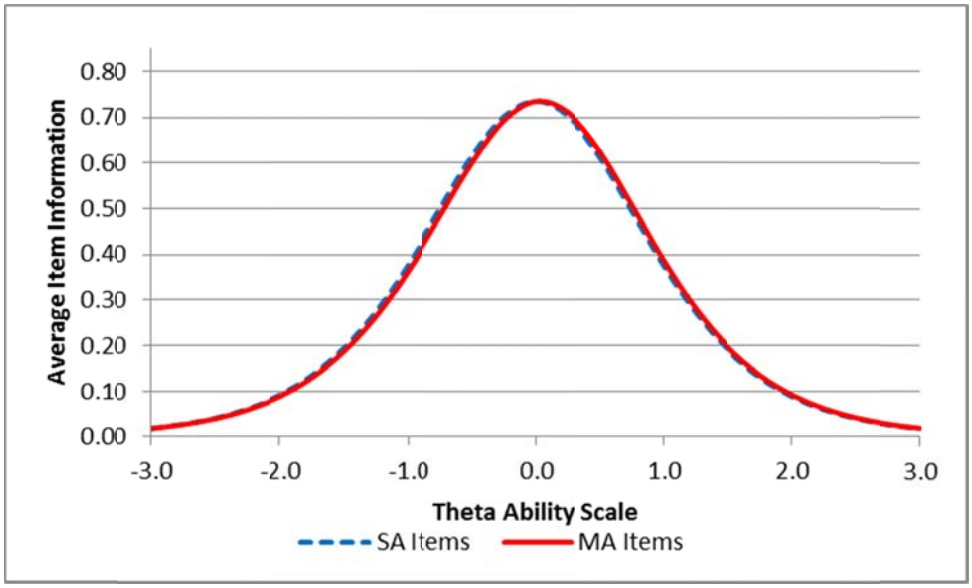

Figure 9. Average item information function for Exam 2 using two parameters

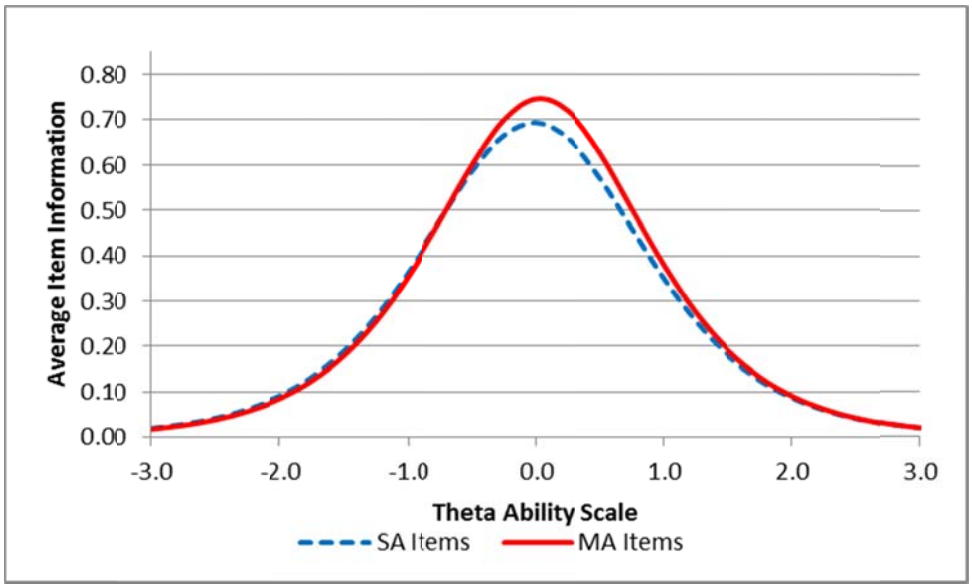

Figure 10. Average item information function for Exam 3 using two parameters

The impact of incorporating the item discrimination estimates was minimal for Exams 1 and 2. Even the third exam only demonstrated a small difference in the performance of SA and MA items. As seen in Figure 10, the average item information (at Theta = 0) for MA items is about .05 higher than the average for SA items.

Partial Credit Versus Dichotomous Scoring of MA Items

Two Rasch analyses were run on each exam. One analysis was performed on the resulting data from using a dichotomous scoring model, and the second analysis was performed on the resulting data from using a partial credit scoring model. Before comparing the resulting reliability estimates, the fit statistics (shown in Table 9) were evaluated to verify that the data adequately fits the models. The infit and outfit statistics are both very good for both models for all three exams. Based on these results, the data fit for both models are sufficient to consider and to compare the reliability estimates for both models.

The reliability coefficients displayed in Table 10 are estimates of the reliability of the examinees' scores obtained from the three exams. As error decreases, reliability increases, with the maximum of 1.00 indicating no random measurement error.

Table 10 displays the two reliability measures for all three exams. The table shows that the use of the Partial Credit Model did not improve the person reliability for any of the exams, and only slightly improved Cronbach's Alpha for Exam 1. Both person reliability and test reliability were slightly lower for Exams 2 and 3 using the Partial Credit Model.

Table 9

Item Statistics for the Dichotomous and Partial Credit Rasch Model

					Infit		Outfit	
Exam	Statistic	Type of Scoring	Measure	SE[a]	MNSQ[b]	ZSTD[c]	MNSQ	ZSTD
1	Mean	Dichotomous	.05	.02	1.00	.0	1.00	.0
		Partial Credit	.22	.03	1.04	.2	1.03	.0
	SD	Dichotomous	.08	.01	.11	.9	.32	.9
		Partial Credit	.12	.02	.18	.9	.40	1.0
	Max.	Dichotomous	.39	.09	1.38	3.4	6.28	4.2
		Partial Credit	.98	.24	2.18	4.0	5.07	4.6
	Min.	Dichotomous	-.22	.02	.69	-2.4	.08	-2.1
		Partial Credit	-.10	.02	.50	-2.3	.05	-1.9
2	Mean	Dichotomous	.18	.04	1.00	.1	1.09	.1
		Partial Credit	.36	.07	1.09	.2	1.17	.2
	SD	Dichotomous	.11	.02	.10	.6	.63	.8
		Partial Credit	.17	.05	.32	.7	.94	.9
	Max.	Dichotomous	.36	.08	1.46	3.7	8.81	4.2
		Partial Credit	.76	.22	2.01	3.5	9.90	3.6
	Min.	Dichotomous	-.13	.03	.68	-2.9	.21	-2.7
		Partial Credit	-.06	.03	.43	-2.3	.13	-1.9
3	Mean	Dichotomous	.08	.01	1.00	.1	.99	.0
		Partial Credit	.16	.02	1.01	.1	.99	.0
	SD	Dichotomous	.04	.00	.07	1.0	.16	1.0
		Partial Credit	.04	.00	.08	.9	.17	1.0
	Max.	Dichotomous	.19	.02	1.25	3.8	1.70	4.0
		Partial Credit	.29	.02	1.28	3.6	2.05	6.3
	Min.	Dichotomous	-.01	.01	.79	-3.2	.62	-2.4
		Partial Credit	.06	.02	.79	-2.6	.59	-2.1

Note. [a]SE= Standard Error, [b]MNSQ = Mean Square, [c]ZSTD = Standardized Z

Table 10.

Reliability Estimates for Dichotomous and Partial Credit Scoring Models

	Rasch Person Reliability			Cronbach Alpha		
	Scoring Method			Scoring Method		
Exam	Dichotomous	Partial Credit	Difference	Dichotomous	Partial Credit	Difference
1	.90	.88	.02	.91	.95	-.04
2	.80	.70	.10	.80	.71	.09
3	.83	.81	.02	.83	.82	.01
Mean	.84	.80	.05	.85	.83	.02

The test information functions were first calculated using both a one-parameter model (Rasch) and a two-parameter model. The Rasch model only uses the difficulty estimates, while the two-parameter model incorporates both item difficulty and item discrimination.

Figure 11 displays the plot of the Rasch test information curves for Form A and Form B for Exam 1 for both the dichotomous model and the Partial Credit Model. As witnessed by the graphs, all four lines are nearly on top of each other, indicating good balance between forms and no difference in the test information between the two scoring models.

Figure 12 displays the plot of the Rasch test information curves for Form A and Form B of Exam 2 for both the dichotomous model and the Partial Credit Model. This exam also exhibits very good balance between the forms and very similar test information functions for the two scoring models.

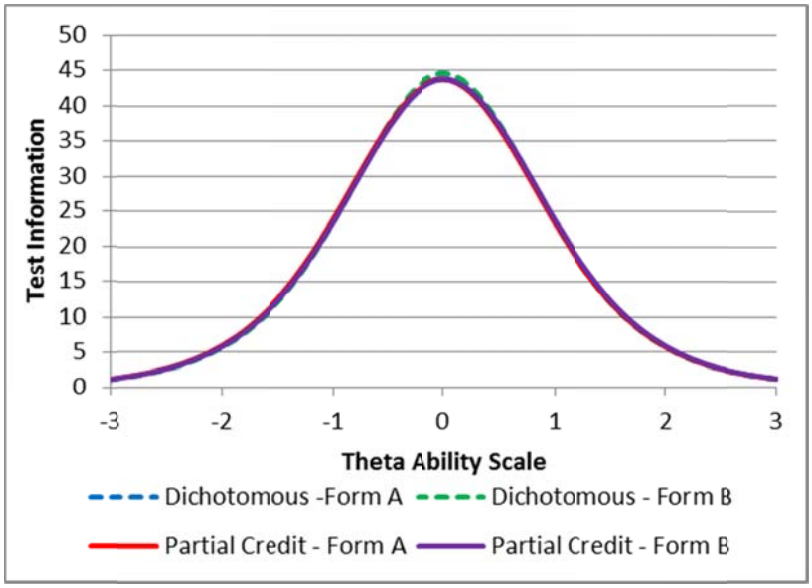

Figure 11. Test information by form and scoring model for Exam 1 using only Rasch

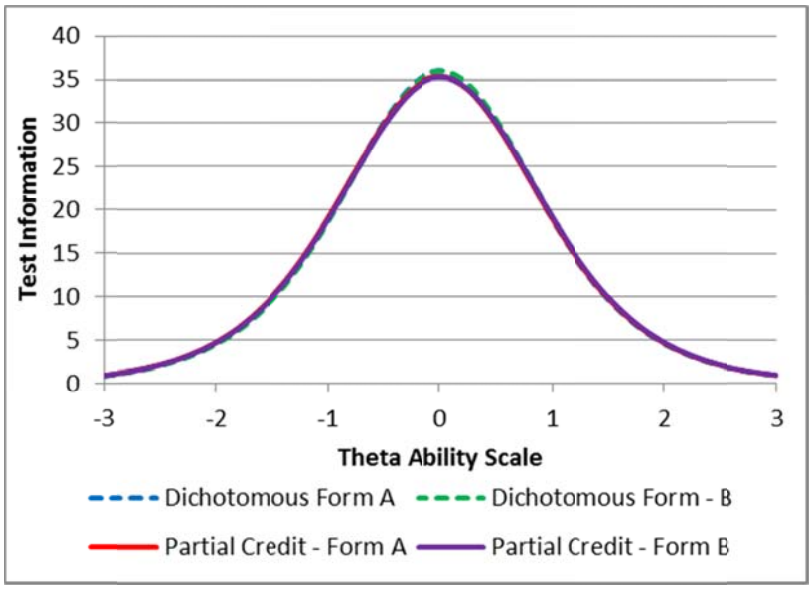

Figure 12. Test information by form and scoring model for Exam 2 using only Rasch

Exam 3 only had a single form, therefore, Figure 13 displays the plots of the test information for the dichotomous model and the Partial Credit Model for one form. The results are similar to the other two exams, with no significant difference between the information for the two models.

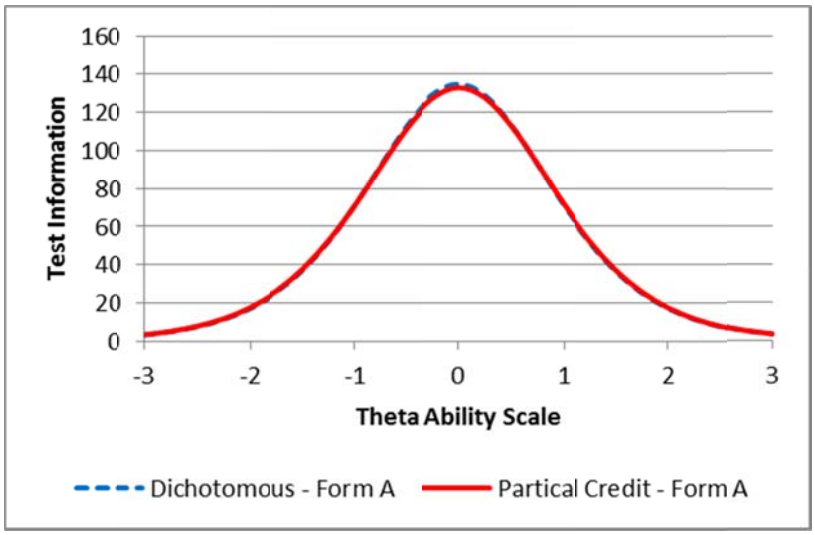

Figure 13. Test information by form and scoring model for Exam 3 using only Rasch

The test information functions were re-calculated using the two parameter test information formula. The two parameter formula incorporates both the item difficulty and discrimination in the calculations.

The purpose for using the two parameter model to estimate the test information functions was to determine if it would result in more of a noticeable difference in the two scoring models. However, the inclusion of the item discrimination into the calculations only resulted in enough of a shift in the plots as to be able to tell that there were different lines being plotted instead of what previously appeared to be only one line. The plots in Figures 14, 15 and 16 demonstrate that the test information functions for all three exams are not impacted by the application of this partial credit scoring model.

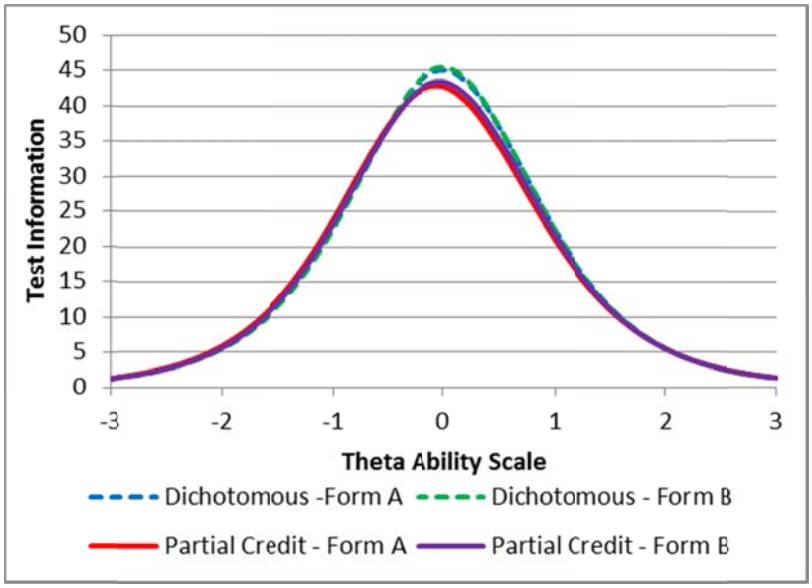

Figure 14. Test information by form and scoring model for Exam 1 using two parameters

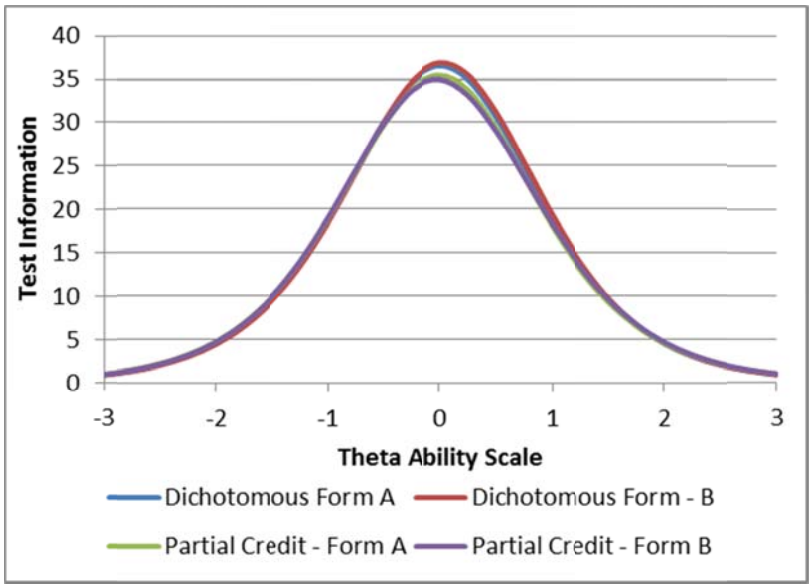

Figure 15. Test information by form and scoring model for Exam 2 using two parameters

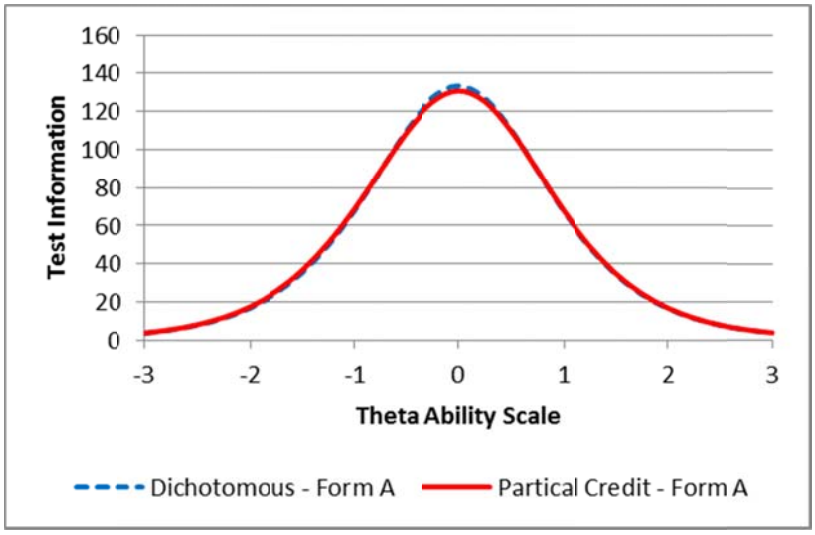

Figure 16. Test information by form and scoring model for Exam 3 using two parameters

Figures 17 through 19 display the frequency distributions of the raw scores for both the dichotomous and the Partial Credit Model for the three exams. The partial credit scoring did not increase the variance of the scores, it only resulted in an upward shift if the scores.

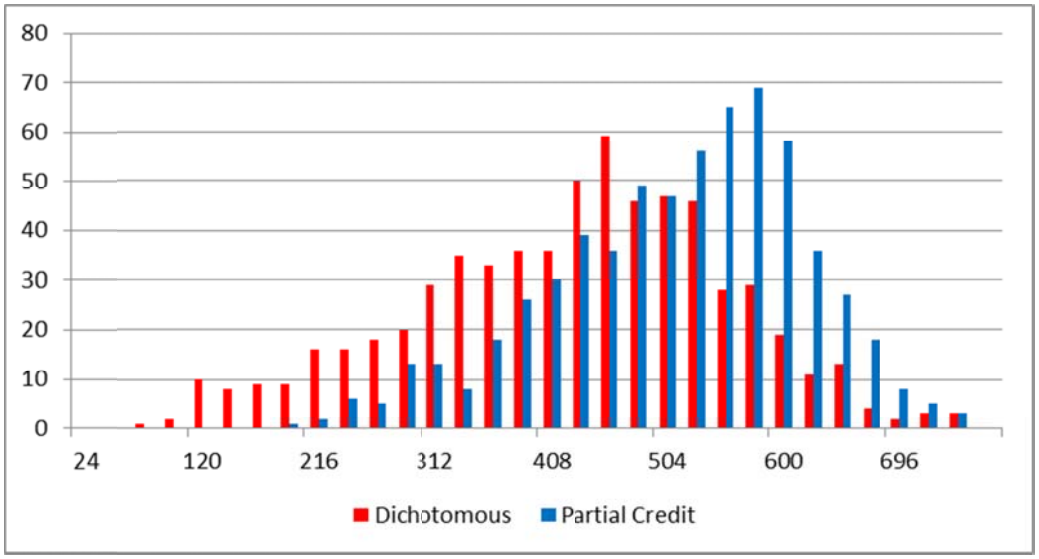

Figure 17. Score distributions for the dichotomous and Partial Credit models for Exam 1

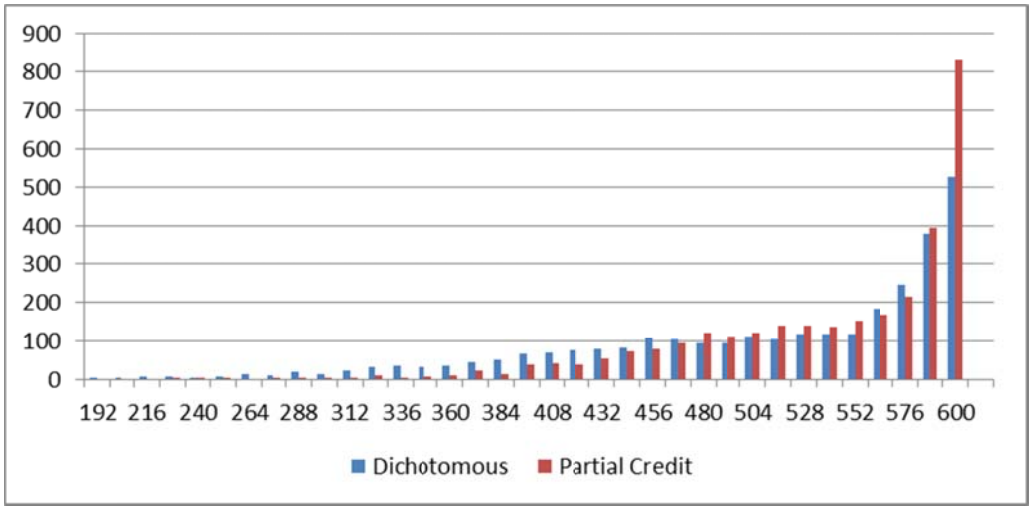

Figure 18. Score distributions for the dichotomous and Partial Credit models for Exam 2

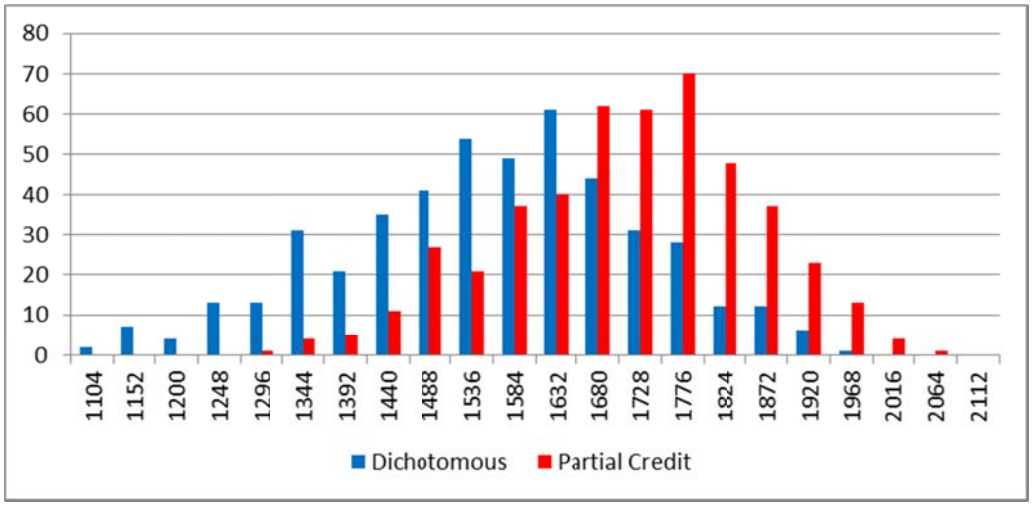

Figure 19. Score distributions for the dichotomous and Partial Credit models for Exam 3

Chapter 5: Discussion

Overall Summary and Reflection

The primary purpose of this research was to determine if MA items with cueing perform sufficiently well to justify their inclusion in high stakes exams. Many published guidelines for writing multiple-choice items (Shrock & Coscarelli, 2007; Burton, et al., 1991; Haladyna & Downing, 1989) recommend against their use. Generally these recommendations have not been based upon the findings of empirical research, or they refer to research on MA items where the examinee was not told how many options were to be selected, but were instructed to "select all that apply." However, the results of this study indicate that when examinees are told how many options they are to select, MA items perform at least as well as SA items, and in some cases may even perform better than SA items. While many authors and researcher have recommended only one correct answer as a general guideline for writing quality multiple-choice items, it is possible that their counsel has largely been misinterpreted. It is very likely that it was never their intention to suggest that items should only have a single keyed correct answer, but that the intent was to verify that only one answer is correct when there is a single keyed correct answer. In other words, item developers should verify that the incorrect answers really are incorrect and not just less correct than the keyed answer. This interpretation of the "one correct answer" guideline would explain why some researchers would consider no empirical research necessary to validate this guideline (Haladyna & Downing, 1989).

There are often situations where including more than one correct answer would create a more authentic task and present it in a more straight-forward, more efficient, and more direct manner. Yet MA items are often avoided because of a concern that they will not perform adequately. So, they are replaced with alternative items that result in the very performance

degradation the examiners were trying to avoid. For example, asking which option is not correct, rather than asking for the three options that are correct creates a negatively worded stem. Research has shown that such negatively-worded stems are often confusing to examinees and tend to perform poorly (Cassels & Johnstone, 1984). Additionally, forcing multiple answers into a single selected response option sacrifices clarity and focus and introduces other sources of error, which threaten validity.

However, I am not advocating the indiscriminant use of MA items. All MA items reviewed in this study were developed with the understanding that the items would be dichotomously scored. The item writers and subsequent reviewers were instructed that approving the item indicated that in order to satisfy the objective, an examinee would need to know all of the correct answers and that there was very little, if any value in partial knowledge. This perspective is important, because it alters the way in which items are developed.

All items reviewed in this study, both SA and MA, were specifically designed and written to be dichotomously scored. To apply a completely different scoring model to items than they were designed for may provide interesting philosophical information, but the practical implications could have a detrimental impact on the validity of the resulting inferences made about the scores. For the exams reviewed in this research, applying a partial credit model to the MA items designed to be dichotomous did not improve the reliability or the resulting test information. But even if the statistics of the MA items had improved by applying the partial credit model, the validity of the scores could be negatively impacted.

The collective information from each item on an exam combines to help determine the total test reliability and the validity of the pass/fail decisions. If it was determined when an MA item was written that mastery of all parts of the item are necessary to meet minimal competency

requirements, then awarding partial credit for partial knowledge could threaten the validity of the decisions made about the scores.

If it is determined that there is value in partial knowledge of item content by providing information or support for interpreting the test scores, then applying a partial credit model to an MA item would be appropriate.

Interpretation of Findings for Each Research Question

The findings of the study are interpreted in the following paragraphs. The interpretations are reviewed separately for each research question.

Content validity of MA items. The purpose of the first research question was to obtain systematically-collected informed judgments about how well the use of dichotomously-scored MA items assessed the targeted KSAs. The average ratings for Exams 1 and 3 indicated the raters generally agreed that the MA items were used appropriately in the context of the targeted KSAs, with total average ratings of .74 and .86 out of a possible 1.0. The average rating (.56) for Exam 2 was just slightly above the mid-point of the rating scale, indicating there may have been some value in partial knowledge for some of the items. However, this can be explained in part by the fact that this exam is more than five years old and many of the SME comments expressed concerns about quality and accuracy of several of the MA items.

Acceptance rates of SA and MA items. The second research focused on the degree to which MA items met the same statistical standards that are typically used for judging SA items. The items were compared in terms of three standards, which were discriminating power, item difficulty, and item reliability.

Item difficulty acceptance rate. The first comparison was the percentages of items that met accepted standards for item difficulty. The results in Table 5 show that MA items had a 4%

higher acceptance rate across all exams for the minimal standard and a 14% higher acceptance rate for the preferred standard. The reason many of the SA items failed to meet the preferred standard may have been because they were too easy. If so, the increased difficulty for MA items likely contributed to a better overall performance towards meeting these criteria.

Item discrimination acceptance rate. The second comparison was the percentage of items meeting the item discrimination standards. MA items had an 8% higher acceptance rate for the minimal standard and a 21% higher acceptance rate at the preferred standard. However, the exam-by-exam comparison resulted in a similar acceptance rate between the two item types.

Item reliability acceptance rate. The final comparison in this section was the percentage of items meeting the item reliability standards. Since MA items were slightly better for both statistics used to calculate item reliability, it is no surprise to see MA items perform better here too. However, it was somewhat surprising to see how much higher the percentage was for MA items than SA items (20% for the minimum standard and 24% for the preferred standard). Therefore, based on this measure, the quality of MA items were superior to the SA items for these exams.

Conclusion for acceptance rates of SA and MA items. While the differences demonstrated in this comparison may not support declaring MA items to be better than SA items, the differences do provide evidence that MA items are at least as effective as SA items with respect to these acceptance criteria.

Average statistical characteristics of SA and MA items. The third research question focused on the comparison of the average difficulty, discrimination, and item reliability of SA and MA items. The purpose of this comparison was to evaluate the relative difference in the statistical performance of SA and MA items.

Item difficulty. To interpret the item difficulty, it is necessary to understand that the ideal item is one in which all minimally-competent examinees and those who are more than minimally competent would answer correctly and everyone else would answer incorrectly. Therefore, it is neither better nor worse for an item to have either a high or low difficulty without referencing the ability level that corresponds to minimal competency. For example, an item that has a very hard difficulty value of .15 would, by itself, seem to be too hard of an item. However, if that item was intended to be on a graduate school entrance exams and it was administered to high school seniors, some may question if the item isn't too easy.

While it is not likely for such an extreme situation to occur, particularly when item quality decisions are being made about items, it does illustrate the importance of understanding the overall ability or competence of the audience taking the exam relative to the minimal competency required to pass the exam. This is why the minimal and preferred standards for item difficulty are stated as ranges of acceptable values, rather than discrete values. For an item to have an opportunity to discriminate between people with high ability and low ability, an item can neither have everyone answering it correctly nor incorrectly. Therefore, as long as the item difficulty falls safely between everyone getting it right and everyone getting it wrong, the item difficulty should not cause an item to be rejected.

One of the biggest concerns about MA items is not just that they may be a degree or two harder than SA item, but that so few people will get them right that they will not discriminate properly. However, by cueing the examinee for how many correct answers they are supposed to pick, the resulting item difficulties of the MA items in this study were as well within the acceptable range as SA items.

Item discrimination. Regardless of how difficult an item is, it is desirable for high-ability people to have more of a tendency to answer it correctly than low ability people. The degree to which high ability people have more of a tendency to answer an item correctly is defined by the item discrimination. Based on the results of this study, SA and MA items similarly met item discrimination requirements. The combined average item discrimination indices for all SA and all MA were within five one-hundredths of each other. MA items had a .01 higher average discrimination index for Exam 1 and a .03 higher index for Exam 2. SA items had a .01 higher average discrimination index for Exam 3. The combined average for all three exams resulted in a .05 higher discrimination index for MA items. Therefore, there is no reason to believe that an item will statistically perform any better or worse based on which item type is used.

Item reliability. As previously mentioned, the item reliability is a mathematical combination of the item difficulty and the item discrimination. The average item reliabilities were similar for MA items than SA items. The average item reliability for Exam 1 was .17 for both SA and MA items, MA items had a .02 higher average item reliability on Exam 2 and a .01 higher average for Exam 3. The total difference in item reliability across all three combined exams was only .03. Based on this analysis, the item reliability for SA and MA items are both acceptable.

Overall statistical comparison. The SA and MA items in this study demonstrated almost identical statistical performance characteristics across all three exams, as supported by both a Classical Test Theory analysis and a Rasch analysis. Therefore, there is no performance-based reason to avoid using them when indicated by the content.

Effect of using subtests. In order to better separate the performance of SA and MA items, the same Classical statistics previously discussed were calculated by basing the item discrimination and item reliability on the total scores of only the items within each item's own item type. So, the item discrimination and item reliability for all SA items were based on only the total scores of all of the other SA items, while the item discrimination and item reliability for all MA items were based only on the total scores of all of the MA items. However, because the performance of SA and MA turned out to be so similar, the use of subtests only resulted in an overall 1% difference from the calculations using total scores.

Conclusion for average statistical characteristics of SA and MA items. The comparison of SA and MA items for research question 3 was even closer than for research question 2. While MA items continued to demonstrate slightly higher overall performance, the difference was too small to declare that MA items are better than SA items. However, these results provide strong evidence that MA items perform at least equal to SA items.

Partial credit versus dichotomous scoring of MA items. The fourth research question investigated the impact of awarding partial credit for MA items compared to dichotomous scoring. The test information functions generated for the exams scored dichotomously were compared to the test information functions that resulted when a partial credit scoring model was used.

Model fit. The fit statistics shown in Table 10 indicate that both models were a good fit to the data. Therefore, the resulting reliability estimates in Table 11 are good estimations for each model.

The results indicate that there is no increase in reliability by using a partial credit model over a dichotomous model. This was surprising because I anticipated that awarding partial credit

for the MA items would result in increased variance, which would provide better reliability estimates. However, by examining the frequency distributions of the scores for each exam in Figures 14 through 16, it became evident that the scores for the Partial Credit Model did not result in a greater variance, or spread in the scores. The result was only a shift upward in all of the scores. Even the plot patterns were very recognizable between the dichotomous models and the Partial Credit Model.

This lack of increased variance and resulting failure to improve the reliabilities were likely influenced by the fact that all three of these exams were specifically designed to be dichotomously scored to begin with. Therefore, the correct answers for the MA items are not always of equal difficulty or importance. The scoring model used for this analysis treated each option within an item equally by granting the same number of points per correct answer. It is very common for MA items to have one option that is extremely easy compared to the other correct answers, sometimes with 100% of examinees selecting it. This provides an increase in scores with very little contribution to the information about the examinees' true ability level. Additionally, items with three and four correct answers rarely have as many incorrect answers as correct answers, which means examinees were awarded points without knowing anything, thus lowering both reliabilities. These difficulties more than offset whatever gains might have otherwise been realized by using a partial credit model. However, for these exams that were designed to be dichotomously scored, a partial credit model does not provide any added benefit with respect to reliability.

Test information. The final comparison of the two scoring models was done by comparing the test information functions for each of the exams using the two scoring models. The test information functions were calculated using both a one-parameter (Rasch) formula and a

two parameter formula. The impact of including the second parameter (item discrimination) for the second calculations was insignificant. The comparison of the plots of the test information functions for the two scoring models indicate that the application of a partial credit model to the MA items does not increase the item information for these exams. This confirms the findings by Hsu et al. (1984) where six different partial credit models were compared. While they found small increases in discrimination and reliability, they declared that the gains did not "justify the additional trouble involved in formula scoring" (p. 158). Based on the comparison of the test information functions, there is no reason to believe that an exam which is properly designed and constructed to be dichotomously scored will benefit from applying a partial credit model.

Limitations

This study compared SA items that were written for objectives that could appropriately be measured by a single answer to MA items that were written for objectives that required multiple parts to be tested. This study was not able to compare MA and SA items that were both designed to measure the same KSAs, which would have been more informative. It would have provided a better opportunity to see the impact MA items can have on exam reliability by reducing construct irrelevant variance that enters into items that are forced to have only one answer when more than one is needed.

Another limitation was not being able to demonstrate the difference in MA items with cueing, which was used by all three of these exams, to MA items without cueing. This would have provided a better tie-back to earlier research sited in this literature review that recommended against using MA items because they are too difficult. Including that in this research would have provided the opportunity to show the impact of cueing.

Conclusions

Considering the diversity of content in the three exams used in this study, it is remarkable how consistently well MA items performed relative to SA items. Two of the exams had very strong overall statistical characteristics, with one of those two being an older exam with some outdated and some inaccurate item content. The third exam was a professional licensure exam that tested less well-defined content that resulted in a considerably lower statistical performance. Yet even with these vast differences, MA items consistently demonstrated a performance at least equal to, and perhaps even a little stronger than SA items.

Test sponsors who have refrained from using MA items out of concern about the difficulty or discrimination performance, can now benefit from the advantages that can be realized through the use of these items. Potential advantages include improved content coverage, reduced construct irrelevant variance, improved efficiency, fewer test items, and shorter test times. Hopefully, test sponsors will view the findings of this study as providing support for the use of MA items and as a rationale for questioning the traditional advice which advocates against their use.

Recommendations

While this study answered a number of important questions surrounding the usefulness and performance of MA items, it raised a number of interesting questions that should be investigated in future research. First, the exams used in this research were all operational exams. Therefore, all items have been previously vetted via a pretesting methodology to eliminate the worst performing items. It would be interesting to compare how raw items perform in a pretest environment to determine if SA or MA items have a higher reject rate.

My recommendation is to use MA items in place of SA items for dichotomously scored exams whenever the content is such that knowledge of all correct answers is necessary to satisfy the testing objective being measured. MA items with cueing have been used by several testing companies over the years. While no formal study has been reported, these items have been successfully utilized in many exams and continue to demonstrate strong statistical performances. Often, upon review of the item statistics for poorly performing MA items, it is often discovered that they were inappropriately utilized for testing objectives where they were not justified. Therefore, the key to the successful use of this item type is to ensure that items of this kind are appropriate for the designated KSA and scoring model.

While the exams used in this study were specifically designed for the professional licensure and certification industry, these results are not limited to those fields. These exams follow the same national standards for test design and development that are used for all testing programs that require high validity and reliability. The comparisons made in this study were performed to verify the same level of quality with MS items as testing experts are accustomed to with SA items. So, the results of this study can and should be applied to any testing program that already utilizes SA items.

A very useful study would be to design an experimental study that would include the comparison of MA items without cueing to the same item with cueing to quantify the impact of cueing on item performance. In addition, it would be very interesting to write both SA and MA items to the same objectives in order to directly compare the statistical characteristics of SA and MA items and to evaluate the impact of item type on validity.

Future research focused on MA items should include using the 3-PL IRT model to assess the effects of including multiple correct answers on examinee guessing behavior. However, such

a study would need to have large sample sizes for each test than were available for the current study.

References

Albanese, M. A. (1982). Multiple-choice items with combinations of correct responses: A further look at the type K format. *Evaluation & the Health Professions, 5*, 218-228.

Albanese, M. A. (1993). Type K and other complex multiple-choice items: An analysis of research and item properties. *Educational Measurement: Issues and Practice, 12*(1), 28-33.

Albanese, M. A., & Sabers, D. L. (1978). *Multiple response vs. multiple true-false scoring: A comparison of reliability and validity.* Presented at the annual meeting of the National Council on Measurement in Education, Toronto, Ontario.

Burton, S. J., Sudweeks, R. R., Merrill, P. F., & Wood, B. (1991). *How to prepare better multiple-choice test items: Guidelines for university faculty.* Provo, UT: Brigham Young University Testing Services and the Department of Instructional Science. Retrieved December 28, 2009, from http://testing.byu.edu/info/handbooks/betteritems.pdf.

Carson, G. J. (1980). Test results depend on response format. *OED Newsletter 6.8*.

Case, S. M., & Downing, S.M. (1989). *Performance of various multiple-choice item types on medical specialty examinations: Types A, B, C, K, and X.* Philadelphia: National Board of Medical Examiners.

Cassels, J. T., & Johnstone, A. H. (1984). The effect of language on student performance on multiple-choice tests in chemistry. *Journal of Chemical Education*, *61*, 613-615.

Cronbach, L. J. (1941). An experimental comparison of the multiple true-false and multiple multiple-choice tests. *Journal of Educational Psychology*, *32*, 533-543.

Downing, S.M. (1992). True-false, alternative choice, and multiple-choice items. *Educational Measurement: Issues and Practice, 11*(3), 27-30.

Duncan, G. T., & Milton, E. O. (1978). Multiple-Answer Multiple-Choice Test Items: Responding and Scoring Through Bayes and Minimax Strategies. *Psychometrika*, 43, 43-57.

Dressel, P. L., & Schmid, J. (1953). Some modifications of the multiple-choice item. *Educational and Psychological Measurement, 13*, 574-595.

Frary, R. B. (1989). Partial-credit scoring methods for multiple-choice tests. *Applied Measurement in Education, 2*, 79-96.

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

Foster, D. F. (1999, September). *Better items.* Paper presented at the 1999 Sylvan Prometric Results Conference, Rancho Mirage, CA.

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

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

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

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

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

Harasym, P. H., Norris, D. A., & Lorscheider, F. L. (1980). Evaluating student multiple-choice responses: effect of coded and free formats. *Evaluation and the Health Professions, 3*(1), 63-84.

Hawkes, H. E., Lindquist, E. F., & Mann, C. R. (1936). *The construction and use of achievement examinations: A manual for secondary school teachers.* Boston: Houghton Mifflin.

Hsu, T. C., Moss, P. A., & Khampalikit, C. (1984). The merits of multiple-answer items as evaluated by using six scoring formulas. *Journal of Experimental Education, 52,* 152-158.

Huntly, R. M. & Plake, B. S. (1984). *An investigation of multiple-response-option items: Item performance and processing demands.* Paper presented at the meeting of the National Council on Measurement in Education, New Orleans.

LaDuca, A., Downing, S. M, & Henzel, T. R. (1995). Systematic item writing and test construction. In J. C. Impara (Ed.), *Licensure Testing: Purposes, Procedures, and Practices* (pp. 117-148). Lincoln, NE: Buros Instistute of Mental Measurements.

Kubiszyn, T., & Borich, G. D. (1987). *Educational testing and measurement: Classroom application and practice* (2nd ed.). Glenview, IL: Scott, Foresman.

Pomplun, M. & Omar, M.H. (1997). Mulitple-mark items: An alternative objective item format? *Educational and Psychological Measurement*, *57*, 949-962.

Prometric (2004). *Developing certification test items – exam development training.* Unpublished manuscript.

Shrock, S. A., & Coscarelli, W. C. (2007). *Criterion-referenced test development* (3rd ed.). San Francisco, CA: Pfeiffer.

Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 329-347). Mahwah, NJ: Erlbaum

Willson, V. L. (1982). Maximizing reliability in multiple-choice questions. *Educational and Psychological Measurement, 42*, 69-72

Appendix A: Item Classical Statistics for Exam 1

Item	Correct	Difficulty	Total Scores		Subtest Scores	
			$r_{i\text{-}total}$	$IR_{i\text{-}total}$	$r_{i\text{-}sub}$	$IR_{i\text{-}sub}$
1	1	.536	.309	.154	.352	.175
2	1	.431	.311	.154	.357	.177
3	1	.661	.381	.180	.350	.166
4	2	.443	.307	.152	.315	.157
5	2	.579	.302	.149	.304	.150
6	2	.722	.294	.132	.313	.140
7	1	.760	.501	.214	.526	.225
8	2	.467	.290	.144	.327	.163
9	2	.385	.381	.185	.409	.199
10	2	.677	.405	.189	.400	.187
11	2	.205	.282	.114	.305	.123
12	1	.412	.317	.156	.354	.174
13	1	.586	.197	.097	.247	.122
14	2	.216	.355	.146	.372	.153
15	1	.697	.374	.172	.381	.175
16	2	.584	.241	.119	.246	.121
17	1	.648	.443	.212	.431	.206
18	3	.213	.309	.126	.325	.133
19	2	.254	.292	.127	.322	.140
20	2	.734	.363	.161	.403	.178
21	1	.784	.468	.193	.450	.185
22	2	.482	.285	.142	.321	.160
23	1	.299	.322	.148	.330	.151
24	2	.629	.501	.242	.509	.246
25	1	.280	.368	.165	.421	.189
26	1	.592	.291	.143	.348	.171
27	2	.350	.492	.235	.482	.230
28	3	.467	.416	.208	.426	.213
29	1	.720	.457	.205	.462	.207
30	1	.638	.434	.209	.463	.222
31	2	.250	.330	.143	.345	.149
32	2	.819	.378	.145	.381	.146
33	1	.665	.390	.184	.405	.191
34	1	.717	.466	.210	.494	.223
35	1	.916	.376	.104	.432	.120
36	3	.488	.290	.145	.325	.162
37	3	.296	.433	.197	.447	.204
38	2	.726	.389	.174	.401	.179

Item	Correct	Difficulty	Total Scores		Subtest Scores	
			$r_{i\text{-}total}$	$IR_{i\text{-}total}$	$r_{i\text{-}sub}$	$IR_{i\text{-}sub}$
39	2	.359	.441	.212	.466	.224
40	2	.611	.468	.228	.489	.239
41	2	.799	.466	.187	.485	.194
42	2	.457	.442	.220	.447	.223
43	1	.479	.281	.140	.317	.158
44	1	.595	.267	.131	.294	.144
45	1	.647	.337	.161	.371	.178
46	1	.664	.499	.236	.521	.246
47	3	.099	.212	.063	.239	.071
48	1	.668	.387	.182	.424	.200
49	1	.536	.484	.241	.493	.246
50	1	.507	.539	.270	.570	.285
51	2	.407	.367	.180	.349	.172
52	1	.919	.344	.094	.356	.097
53	1	.716	.385	.174	.431	.195
54	1	.557	.312	.155	.398	.198
55	2	.609	.450	.220	.402	.196
56	1	.813	.422	.165	.481	.188
57	1	.870	.503	.169	.525	.177
58	1	.743	.388	.170	.423	.185
59	1	.775	.393	.164	.439	.183
60	1	.716	.317	.143	.388	.175
61	1	.625	.466	.225	.538	.261
62	1	.653	.352	.168	.404	.192
63	2	.627	.408	.197	.400	.194
64	1	.584	.331	.163	.374	.184
65	1	.697	.349	.160	.374	.172
66	1	.740	.280	.123	.357	.157
67	1	.716	.402	.181	.445	.201
68	1	.551	.469	.233	.500	.249
69	2	.711	.325	.148	.328	.149
70	1	.635	.354	.170	.427	.206
71	3	.635	.476	.229	.480	.231
72	2	.466	.491	.245	.500	.250
73	3	.803	.326	.130	.346	.138
74	1	.722	.396	.178	.408	.183
75	1	.599	.413	.203	.432	.212
76	1	.895	.442	.135	.461	.141
77	1	.855	.513	.180	.489	.172
78	2	.878	.481	.157	.468	.153

Item	Correct	Difficulty	Total Scores		Subtest Scores	
			$r_{i\text{-}total}$	$IR_{i\text{-}total}$	$r_{i\text{-}sub}$	$IR_{i\text{-}sub}$
79	2	.684	.393	.182	.396	.184
80	2	.527	.461	.230	.465	.232
81	2	.280	.395	.177	.446	.200
82	2	.641	.332	.159	.365	.175
83	1	.587	.349	.172	.335	.165
84	2	.569	.376	.186	.381	.189
85	1	.491	.421	.210	.468	.234
86	4	.513	.450	.225	.483	.241
87	2	.849	.189	.068	.212	.076
88	2	.659	.344	.163	.351	.167
89	3	.467	.317	.158	.312	.156
90	2	.710	.405	.184	.429	.195
91	2	.443	.309	.154	.358	.178
92	3	.553	.466	.232	.499	.248
93	2	.645	.320	.153	.347	.166
94	2	.728	.406	.181	.410	.183
95	2	.641	.359	.172	.388	.186
96	2	.644	.422	.202	.451	.216
97	3	.657	.455	.216	.473	.224
98	3	.880	.378	.123	.364	.118
99	3	.487	.347	.173	.391	.196
100	2	.638	.364	.175	.412	.198
101	3	.389	.407	.198	.452	.220
102	2	.566	.492	.244	.504	.250
103	3	.677	.326	.152	.373	.174
104	3	.587	.358	.176	.368	.181
105	4	.438	.471	.233	.472	.234
106	3	.653	.459	.219	.466	.222
107	1	.751	.328	.142	.348	.150
108	2	.253	.246	.107	.282	.122
109	1	.880	.277	.090	.311	.101
110	3	.666	.312	.147	.324	.153
111	1	.806	.282	.112	.305	.121

Note. $r_{i\text{-}total}$ = item discrimination based on total scores, $IR_{i\text{-}total}$ = item reliability based on total scores, $r_{i\text{-}sub}$ = item discrimination based on subtest totals, $IR_{i\text{-}sub}$ = item reliability based on subtest scores.

Appendix B: Item Classical Statistics for Exam 2

Item	Correct	Difficulty	Total Scores		Subtest Scores	
			$r_{i\text{-}total}$	$IR_{i\text{-}total}$	$r_{i\text{-}sub}$	$IR_{i\text{-}sub}$
1	1	.667	.599	.282	.643	.303
2	2	.933	.369	.092	.375	.094
3	1	.973	.321	.052	.336	.055
4	1	.921	.372	.100	.391	.105
5	1	.913	.384	.108	.392	.111
6	3	.865	.369	.126	.395	.135
7	1	.937	.288	.070	.301	.073
8	2	.899	.244	.073	.273	.082
9	2	.969	.222	.039	.243	.042
10	1	.939	.348	.083	.353	.084
11	2	.895	.394	.121	.416	.127
12	1	.963	.272	.052	.284	.054
13	2	.928	.439	.114	.444	.115
14	2	.942	.334	.078	.339	.079
15	1	.875	.456	.151	.477	.158
16	1	.781	.481	.199	.515	.213
17	3	.830	.493	.185	.518	.194
18	2	.803	.505	.201	.510	.203
19	3	.790	.450	.183	.494	.201
20	1	.845	.403	.146	.452	.164
21	2	.755	.552	.238	.553	.238
22	2	.913	.353	.100	.385	.109
23	1	.858	.496	.173	.491	.171
24	2	.889	.456	.143	.483	.152
25	1	.857	.342	.120	.360	.126
26	3	.842	.452	.165	.469	.171
27	3	.888	.405	.128	.422	.133
28	3	.670	.616	.290	.618	.291
29	1	.765	.532	.226	.563	.238
30	3	.733	.562	.249	.552	.244
31	3	.838	.466	.172	.475	.175
32	3	.648	.558	.267	.560	.267
33	2	.907	.401	.117	.420	.122
34	1	.727	.615	.274	.641	.286
35	2	.875	.502	.166	.525	.173
36	3	.786	.499	.205	.521	.214
37	1	.919	.385	.105	.371	.101
38	2	.933	.358	.090	.346	.087

Item	Correct	Difficulty	Total Scores		Subtest Scores	
			$r_{i\text{-}total}$	$IR_{i\text{-}total}$	$r_{i\text{-}sub}$	$IR_{i\text{-}sub}$
39	1	.838	.349	.129	.355	.131
40	1	.903	.370	.110	.382	.113
41	1	.926	.310	.081	.311	.081
42	1	.906	.389	.113	.404	.118
43	1	.855	.480	.169	.482	.170
44	2	.945	.191	.043	.209	.048
45	1	.852	.438	.156	.451	.160
46	2	.961	.143	.028	.171	.033
47	1	.736	.597	.263	.631	.278
48	3	.766	.553	.234	.580	.246
49	1	.912	.230	.065	.264	.075
50	1	.827	.500	.189	.520	.197
51	1	.851	.474	.169	.472	.168
52	1	.936	.342	.084	.328	.081
53	2	.890	.408	.128	.415	.130
54	1	.734	.532	.235	.560	.247
55	1	.848	.449	.161	.461	.165
56	1	.913	.390	.110	.400	.113
57	2	.881	.427	.138	.430	.139
58	3	.839	.532	.195	.553	.203
59	1	.921	.376	.101	.391	.105
60	1	.896	.340	.104	.363	.111
61	1	.942	.334	.078	.353	.082
62	1	.871	.386	.129	.436	.146
63	3	.897	.317	.096	.339	.103
64	3	.915	.315	.088	.318	.089
65	1	.939	.309	.074	.320	.077
66	1	.872	.446	.149	.451	.151
67	1	.958	.298	.060	.297	.059
68	3	.922	.413	.110	.437	.117
69	1	.864	.492	.169	.541	.185
70	1	.824	.563	.214	.560	.213
71	1	.920	.242	.065	.231	.062
72	1	.880	.311	.101	.351	.114
73	2	.724	.616	.275	.616	.275
74	2	.899	.394	.119	.416	.125
75	2	.947	.276	.062	.291	.065
76	3	.787	.466	.191	.509	.208
77	3	.758	.525	.225	.559	.239
78	3	.617	.621	.302	.640	.311

Item	Correct	Difficulty	Total Scores		Subtest Scores	
			$r_{i\text{-}total}$	$IR_{i\text{-}total}$	$r_{i\text{-}sub}$	$IR_{i\text{-}sub}$
79	3	.615	.646	.314	.634	.308
80	3	.836	.410	.152	.445	.165
81	2	.792	.463	.188	.473	.192
82	2	.810	.485	.190	.523	.205
83	1	.778	.503	.209	.524	.218
84	1	.869	.400	.135	.402	.135
85	2	.845	.423	.153	.446	.162
86	1	.958	.277	.056	.305	.062
87	1	.764	.560	.238	.589	.250
88	2	.718	.585	.263	.593	.267

Note. $r_{i\text{-}total}$ = item discrimination based on total scores, $IR_{i\text{-}total}$ = item reliability based on total scores, $r_{i\text{-}sub}$ = item discrimination based on subtest totals, $IR_{i\text{-}sub}$ = item reliability based on subtest scores.

Appendix C: Item Classical Statistics for Exam 3

Item	Correct	Difficulty	Total Scores		Subtest Scores	
			$r_{i\text{-}total}$	$IR_{i\text{-}total}$	$r_{i\text{-}sub}$	$IR_{i\text{-}sub}$
1	1	.594	.241	.119	.242	.119
2	1	.525	.052	.026	.069	.034
3	1	.987	.138	.016	.163	.018
4	1	.746	.071	.031	.086	.037
5	2	.675	.233	.109	.267	.125
6	1	.495	.097	.048	.126	.063
7	1	.776	.129	.054	.147	.061
8	3	.673	.018	.009	.077	.036
9	3	.772	.173	.072	.152	.064
10	1	.662	.185	.087	.200	.095
11	3	.447	.226	.113	.314	.156
12	1	.523	.295	.147	.332	.166
13	1	.596	.299	.147	.285	.140
14	2	.611	.164	.080	.186	.091
15	1	.538	.229	.114	.220	.109
16	2	.103	.007	.002	.051	.015
17	2	.641	.261	.125	.259	.124
18	1	.867	.174	.059	.174	.059
19	2	.518	.095	.048	.144	.072
20	1	.892	.292	.090	.283	.088
21	3	.484	.219	.109	.260	.130
22	2	.669	.155	.073	.214	.101
23	1	.269	.060	.027	.085	.038
24	1	.434	.136	.067	.164	.081
25	1	.751	.139	.060	.148	.064
26	1	.888	.068	.021	.089	.028
27	2	.955	.218	.045	.209	.043
28	1	.735	.219	.097	.239	.106
29	1	.441	.106	.053	.124	.062
30	1	.759	.216	.092	.225	.096
31	1	.720	.253	.114	.279	.125
32	2	.985	.068	.008	.064	.008
33	1	.766	.010	.004	.034	.014
34	3	.718	.235	.106	.256	.115
35	1	.465	.071	.035	.065	.032
36	1	.759	.261	.112	.259	.111
37	1	.718	.202	.091	.210	.094
38	1	.533	.205	.102	.221	.110

Item	Correct	Difficulty	Total Scores		Subtest Scores	
			$r_{i\text{-}total}$	$IR_{i\text{-}total}$	$r_{i\text{-}sub}$	$IR_{i\text{-}sub}$
39	1	.783	.318	.131	.315	.130
40	1	.946	.143	.032	.145	.033
41	1	.471	.114	.057	.136	.068
42	1	.796	.223	.090	.226	.091
43	1	.899	.182	.055	.144	.043
44	1	.772	.118	.050	.127	.053
45	3	.602	.340	.167	.373	.182
46	1	.927	.266	.069	.275	.072
47	3	.318	.119	.055	.144	.067
48	1	.929	.065	.017	.054	.014
49	1	.731	.229	.102	.233	.103
50	1	.914	.238	.067	.241	.068
51	1	.828	.208	.078	.216	.081
52	3	.619	.271	.131	.259	.126
53	1	.785	.191	.078	.214	.088
54	1	.226	.075	.032	.092	.039
55	1	.232	.086	.036	.091	.038
56	3	.662	.179	.084	.232	.110
57	3	.804	.117	.047	.124	.049
58	1	.865	.179	.061	.174	.059
59	1	.882	.157	.051	.143	.046
60	1	.516	.115	.057	.141	.070
61	1	.716	.250	.113	.262	.118
62	1	.768	.249	.105	.275	.116
63	3	.667	.168	.079	.231	.109
64	2	.800	-.017	-.007	.014	.006
65	3	.460	.210	.104	.244	.122
66	1	.632	.274	.132	.281	.136
67	4	.604	.173	.085	.170	.083
68	2	.860	.187	.065	.201	.070
69	2	.682	.296	.138	.330	.154
70	2	.860	.030	.010	.049	.017
71	3	.994	.183	.015	.183	.015
72	1	.905	.117	.034	.104	.030
73	1	.667	.217	.102	.228	.108
74	1	.703	.247	.113	.256	.117
75	2	.277	.115	.051	.181	.081
76	1	.686	.301	.140	.296	.137
77	1	.738	.215	.095	.222	.098
78	3	.443	.296	.147	.362	.180

Item	Correct	Difficulty	Total Scores		Subtest Scores	
			$r_{i\text{-}total}$	$IR_{i\text{-}total}$	$r_{i\text{-}sub}$	$IR_{i\text{-}sub}$
79	1	.594	.214	.105	.219	.108
80	2	.766	.160	.068	.210	.089
81	1	.647	.294	.140	.302	.144
82	1	.776	.266	.111	.257	.107
83	1	.774	.262	.110	.244	.102
84	2	.705	.298	.136	.298	.136
85	3	.815	.119	.046	.143	.056
86	3	.617	.373	.181	.382	.186
87	1	.841	.272	.100	.294	.107
88	1	.624	.098	.047	.124	.060
89	1	.574	.200	.099	.225	.111
90	1	.686	.285	.132	.298	.138
91	1	.774	.211	.088	.208	.087
92	1	.751	.152	.066	.131	.057
93	1	.755	.038	.016	.045	.019
94	1	.789	.036	.015	.051	.021
95	1	.602	.300	.147	.281	.138
96	1	.873	.223	.074	.221	.074
97	1	.811	.164	.064	.159	.062
98	3	.925	.106	.028	.102	.027
99	1	.665	.173	.082	.189	.089
100	1	.587	.201	.099	.195	.096
101	1	.480	.192	.096	.222	.111
102	1	.789	.218	.089	.231	.094
103	1	.557	.080	.040	.131	.065
104	2	.916	.179	.050	.203	.056
105	1	.748	.207	.090	.211	.092
106	1	.849	.197	.071	.165	.059
107	1	.727	.208	.093	.224	.100
108	1	.983	.063	.008	.055	.007
109	1	.847	.292	.105	.313	.113
110	1	.916	.125	.035	.134	.037
111	2	.712	.170	.077	.180	.082
112	1	.458	.257	.128	.242	.121
113	1	.927	.164	.043	.169	.044
114	1	.817	.226	.087	.194	.075
115	3	.576	.269	.133	.293	.145
116	1	.813	.273	.106	.263	.103
117	1	.615	.172	.084	.173	.084
118	1	.665	.175	.083	.203	.096

Item	Correct	Difficulty	Total Scores		Subtest Scores	
			$r_{i\text{-}total}$	$IR_{i\text{-}total}$	$r_{i\text{-}sub}$	$IR_{i\text{-}sub}$
119	3	.295	.125	.057	.182	.083
120	1	.723	.204	.091	.214	.096
121	1	.523	.166	.083	.176	.088
122	1	.817	.237	.091	.242	.094
123	2	.628	.131	.063	.185	.090
124	1	.892	.271	.084	.285	.088
125	1	.923	.147	.039	.155	.042
126	1	.234	-.045	-.019	-.049	-.021
127	3	.327	.148	.070	.189	.089
128	3	.740	.154	.068	.226	.099
129	1	.948	.090	.020	.109	.024
130	1	.559	.133	.066	.129	.064
131	1	.860	.183	.063	.194	.067
132	1	.916	.170	.047	.184	.051
133	1	.557	.079	.039	.104	.052
134	1	.927	.074	.019	.097	.025
135	1	.467	.169	.084	.175	.087
136	2	.920	.173	.047	.171	.046
137	1	.837	.080	.030	.106	.039
138	1	.475	.097	.048	.094	.047
139	1	.525	.160	.080	.162	.081
140	3	.774	.235	.098	.252	.105
141	2	.030	.080	.014	.098	.017
142	2	.602	.339	.166	.346	.169
143	1	.548	.132	.066	.148	.074
144	1	.647	.140	.067	.181	.086
145	1	.583	.170	.084	.174	.086
146	1	.895	-.013	-.004	-.008	-.002
147	2	.658	-.081	-.039	.011	.005
148	1	.574	.168	.083	.192	.095
149	3	.378	.281	.136	.301	.146
150	1	.755	.290	.125	.274	.118
151	1	.628	.148	.071	.141	.068
152	1	.275	.016	.007	.027	.012
153	1	.488	.099	.049	.101	.050
154	3	.766	.257	.109	.226	.096
155	1	.544	.246	.123	.263	.131
156	2	.912	.110	.031	.164	.047
157	1	.892	.143	.044	.136	.042
158	1	.948	.074	.016	.085	.019

Item	Correct	Difficulty	Total Scores		Subtest Scores	
			$r_{i\text{-}total}$	$IR_{i\text{-}total}$	$r_{i\text{-}sub}$	$IR_{i\text{-}sub}$
159	2	.692	.146	.067	.184	.085
160	1	.508	.269	.134	.259	.129
161	3	.411	.245	.120	.296	.146
162	2	.559	.093	.046	.158	.078
163	2	.363	.090	.043	.110	.053
164	1	.415	.229	.113	.223	.110
165	3	.598	.118	.058	.166	.081
166	2	.338	.028	.013	.090	.043
167	1	.335	.165	.078	.156	.074
168	1	.774	.293	.123	.303	.127
169	1	.940	.027	.006	.030	.007
170	1	.899	.071	.021	.096	.029
171	1	.615	.278	.135	.273	.133
172	1	.858	.207	.072	.224	.078
173	3	.828	.271	.102	.278	.105
174	1	.856	.088	.031	.091	.032
175	1	.699	.345	.158	.372	.171
176	1	.746	.257	.112	.266	.116
177	2	.516	.241	.120	.248	.124
178	1	.849	.165	.059	.149	.053
179	3	.778	.113	.047	.171	.071
180	1	.839	.120	.044	.125	.046
181	1	.849	.276	.099	.267	.095
182	1	.886	.287	.091	.280	.089
183	1	.787	.290	.119	.295	.121
184	1	.892	.047	.015	.054	.017
185	3	.553	.244	.121	.247	.123
186	1	.920	.111	.030	.121	.033
187	1	.768	.211	.089	.223	.094
188	3	.439	.331	.164	.341	.169

Note. $r_{i\text{-}total}$ = item discrimination based on total scores, $IR_{i\text{-}total}$ = item reliability based on total scores, $r_{i\text{-}sub}$ = item discrimination based on subtest totals, $IR_{i\text{-}sub}$ = item reliability based on subtest scores.

Appendix D: Item Rasch Statistics for Exam 1

	Dichotomous		Partial Credit	
Item	Rasch Difficulty	Discrimination[a]	Rasch Difficulty	Discrimination
1	.030	0.750	.190	0.520
2	.070	0.640	.240	0.390
3	-.020	1.010	.140	0.790
4	.060	0.710	-.030	1.030
5	.010	0.650	-.250	0.960
6	-.050	0.830	-.200	1.000
7	-.070	1.190	.080	1.130
8	.060	0.580	-.030	0.940
9	.090	0.910	-.010	1.060
10	-.030	1.030	-.150	1.110
11	.180	0.950	.200	0.990
12	.090	0.770	.260	0.550
13	.010	0.420	.170	0.220
14	.170	1.060	.240	1.100
15	-.040	0.990	.120	0.850
16	.010	0.460	-.180	0.860
17	-.020	1.100	.140	0.910
18	.170	1.000	-.060	1.070
19	.150	0.910	.130	1.060
20	-.060	1.010	-.220	1.070
21	-.080	1.140	.070	1.050
22	.050	0.650	-.060	0.930
23	.130	0.890	.300	0.710
24	-.010	1.260	.040	1.050
25	.140	0.970	.310	0.840
26	.010	0.720	.170	0.530
27	.100	1.310	-.050	1.220
28	.050	1.050	-.180	1.060
29	-.050	1.160	.110	1.060
30	-.010	1.110	.150	0.910
31	.150	0.950	.170	0.990
32	-.100	1.040	-.300	1.060
33	-.030	0.980	.130	0.790
34	-.050	1.180	.110	1.110
35	-.190	1.070	-.040	1.040
36	.050	0.620	-.260	1.070
37	.130	1.120	-.070	1.150
38	-.050	1.030	-.120	1.070
39	.100	1.200	.090	1.100

	Dichotomous		Partial Credit	
Item	Rasch Difficulty	Discrimination[a]	Rasch Difficulty	Discrimination
40	.000	1.170	-.050	1.140
41	-.090	1.170	-.180	1.120
42	.060	1.130	.060	1.090
43	.050	0.620	.220	0.400
44	.010	0.600	.170	0.410
45	-.020	0.830	.140	0.650
46	-.020	1.270	.140	1.150
47	.270	0.970	.370	1.000
48	-.030	1.000	.130	0.870
49	.030	1.250	.190	1.040
50	.040	1.470	.210	1.280
51	.080	0.940	-.020	1.050
52	-.190	1.050	-.050	1.020
53	-.050	1.000	.110	0.870
54	.020	0.730	.180	0.510
55	.000	1.180	-.100	1.170
56	-.100	1.100	.060	1.060
57	-.150	1.160	.010	1.160
58	-.060	1.020	.100	0.910
59	-.080	1.050	.070	0.940
60	-.050	0.890	.110	0.720
61	-.010	1.180	.150	1.070
62	-.020	0.940	.140	0.710
63	-.010	0.990	-.110	1.100
64	.010	0.820	.170	0.590
65	-.040	0.930	.120	0.790
66	-.060	0.900	.100	0.780
67	-.050	1.050	.110	0.930
68	.020	1.260	.190	1.030
69	-.040	0.860	-.080	1.010
70	-.010	0.880	.150	0.730
71	-.010	1.190	-.400	1.240
72	.060	1.380	-.080	1.230
73	-.090	0.970	-.370	1.050
74	-.050	1.030	.110	0.880
75	.000	1.100	.160	0.920
76	-.170	1.120	-.020	1.100
77	-.130	1.180	.020	1.160
78	-.150	1.160	-.160	1.080
79	-.030	1.000	-.220	1.090

Item	Dichotomous		Partial Credit	
	Rasch Difficulty	Discrimination[a]	Rasch Difficulty	Discrimination
80	.030	1.240	-.140	1.250
81	.140	1.010	.120	1.010
82	-.020	0.830	-.080	1.000
83	.010	0.840	.170	0.610
84	.020	0.870	-.030	1.070
85	.050	1.130	.210	0.890
86	.040	1.090	-.120	1.190
87	-.120	0.880	-.300	0.970
88	-.020	0.930	-.030	1.030
89	.060	0.690	-.180	1.070
90	-.040	1.060	.000	1.070
91	.060	0.690	.100	0.800
92	.020	1.180	-.300	1.120
93	-.020	0.760	-.170	1.000
94	-.050	1.040	-.180	1.100
95	-.010	0.900	-.090	1.050
96	-.020	1.060	-.050	1.090
97	-.020	1.110	-.260	1.130
98	-.150	1.050	-.450	1.050
99	.050	0.730	-.240	1.110
100	-.010	0.950	-.080	1.050
101	.090	1.090	-.140	1.110
102	.020	1.330	.020	1.160
103	-.030	0.890	-.290	1.080
104	.010	0.830	-.280	1.060
105	.070	1.290	-.150	1.140
106	-.020	1.170	-.290	1.120
107	-.060	0.990	.100	0.850
108	.150	0.770	.170	0.850
109	-.150	0.980	.000	0.890
110	-.020	0.830	-.320	1.050
111	-.100	0.930	.060	0.840

Note. [a]Winsteps estimates discrimination after calculating the Rasch difficulty

Appendix E: Item Rasch Statistics for Exam 2

Item	Dichotomous		Partial Credit	
	Rasch Difficulty	Discrimination[a]	Rasch Difficulty	Discrimination
1	.140	1.170	.290	0.960
2	-.070	1.030	-.130	1.030
3	-.160	1.060	-.030	1.050
4	-.060	1.010	.080	0.950
5	-.050	0.990	.090	0.920
6	.010	0.870	-.290	1.020
7	-.080	0.960	.050	0.890
8	-.030	0.800	-.130	0.930
9	-.140	0.980	-.080	0.980
10	-.080	1.020	.050	0.970
11	-.020	1.010	-.160	1.050
12	-.130	0.990	.000	0.950
13	-.060	1.080	.050	1.050
14	-.090	1.020	-.160	1.030
15	.000	1.050	.140	0.990
16	.070	0.900	.210	0.680
17	.030	1.010	-.010	1.110
18	.050	1.050	-.190	1.140
19	.060	0.800	-.310	1.010
20	.020	0.960	.160	0.850
21	.090	1.110	.100	1.080
22	-.050	0.960	-.090	1.000
23	.010	1.070	.150	0.960
24	-.020	1.060	-.230	1.080
25	.010	0.870	.150	0.700
26	.020	0.960	-.030	1.080
27	-.020	0.980	-.340	1.040
28	.130	1.230	.240	1.770
29	.080	1.090	.230	0.940
30	.100	1.110	-.010	1.150
31	.030	0.970	.040	1.020
32	.150	1.000	-.250	1.130
33	-.040	0.990	-.050	1.010
34	.100	1.210	.250	1.060
35	-.010	1.090	-.160	1.080
36	.060	0.970	.070	1.150
37	-.050	1.010	.090	0.940
38	-.070	1.010	-.110	1.040
39	.030	0.800	.170	0.580

Item	Dichotomous		Partial Credit	
	Rasch Difficulty	Discrimination[a]	Rasch Difficulty	Discrimination
40	-.040	0.940	.100	0.840
41	-.060	0.950	.080	0.860
42	-.040	1.010	.100	0.950
43	.010	1.070	.150	1.000
44	-.090	0.890	-.280	0.980
45	.010	0.990	.150	0.890
46	-.120	0.880	-.280	0.950
47	.090	1.190	.240	1.060
48	.080	1.070	-.020	1.120
49	-.050	0.790	.090	0.630
50	.030	1.090	.170	1.000
51	.020	1.030	.160	0.910
52	-.080	1.010	.060	0.970
53	-.020	0.990	-.200	1.040
54	.100	1.040	.240	0.850
55	.020	1.020	.160	0.920
56	-.040	1.000	.090	0.930
57	-.010	0.990	-.040	1.030
58	.030	1.110	-.030	1.150
59	-.050	1.020	.080	0.960
60	-.030	0.920	.110	0.800
61	-.090	1.010	.040	0.970
62	.000	0.970	.140	0.890
63	-.030	0.880	-.390	1.010
64	-.050	0.940	-.330	1.020
65	-.080	0.990	.050	0.930
66	.000	1.040	.130	0.970
67	-.120	1.020	.020	0.980
68	-.060	1.050	-.330	1.050
69	.000	1.090	.140	1.040
70	.040	1.100	.180	0.980
71	-.050	0.860	.080	0.720
72	-.010	0.880	.130	0.760
73	.100	1.260	.040	1.200
74	-.030	1.000	-.050	1.030
75	-.090	0.960	-.330	1.000
76	.060	0.910	-.060	1.050
77	.080	0.950	-.350	1.090
78	.160	1.190	-.200	1.210
79	.170	1.300	-.250	1.310

	Dichotomous		Partial Credit	
Item	Rasch Difficulty	Discrimination[a]	Rasch Difficulty	Discrimination
80	.030	0.870	-.320	1.030
81	.060	0.910	-.070	1.020
82	.050	0.960	.140	0.940
83	.070	1.000	.210	0.810
84	.000	0.950	.140	0.820
85	.020	0.930	-.070	1.020
86	-.120	0.990	.010	0.960
87	.080	1.170	.230	1.050
88	.110	1.130	.020	1.150

Note. [a]Winsteps estimates discrimination after calculating the Rasch difficulty

Appendix F: Item Rasch Statistics for Exam 3

	Dichotomous		Partial Credit	
Item	Rasch Difficulty	Discrimination[a]	Rasch Difficulty	Discrimination
1	.050	1.110	.130	1.060
2	.070	0.250	.160	0.250
3	-.290	1.010	-.210	1.020
4	-.010	0.910	.070	0.890
5	.020	1.060	-.040	1.020
6	.080	0.440	.170	0.440
7	-.030	0.970	.050	0.960
8	.020	0.740	-.020	0.950
9	-.020	0.990	-.140	1.020
10	.020	0.970	.100	0.950
11	.100	1.080	.220	1.490
12	.070	1.440	.160	1.380
13	.050	1.280	.130	1.200
14	.040	0.910	-.220	1.020
15	.070	1.100	.150	0.980
16	.270	0.960	.200	0.970
17	.030	1.110	-.200	1.070
18	-.080	1.010	.000	1.010
19	.080	0.440	-.120	0.980
20	-.100	1.050	-.020	1.050
21	.090	1.060	-.020	1.060
22	.020	0.940	.040	1.020
23	.170	0.890	.250	0.870
24	.110	0.730	.190	0.700
25	-.010	0.970	.070	0.950
26	-.100	0.980	-.020	0.970
27	-.180	1.030	-.350	1.010
28	-.010	1.030	.070	1.030
29	.100	0.620	.180	0.590
30	-.020	1.020	.060	1.020
31	.000	1.060	.080	1.060
32	-.270	1.000	-.360	1.000
33	-.020	0.880	.060	0.860
34	.000	1.040	-.340	1.030
35	.090	0.390	.180	0.340
36	-.020	1.060	.060	1.050
37	.000	1.010	.080	1.000
38	.070	0.980	.150	0.910
39	-.030	1.090	.050	1.090

Item	Dichotomous		Partial Credit	
	Rasch Difficulty	Discrimination[a]	Rasch Difficulty	Discrimination
40	-.160	1.010	-.080	1.010
41	.090	0.550	.170	0.540
42	-.040	1.030	.040	1.020
43	-.110	1.010	-.030	1.000
44	-.020	0.960	.060	0.950
45	.050	1.370	-.240	1.070
46	-.140	1.040	-.060	1.040
47	.150	0.910	.360	1.020
48	-.140	0.990	-.060	0.980
49	.000	1.030	.080	1.020
50	-.120	1.030	-.040	1.030
51	-.050	1.020	.030	1.010
52	.040	1.160	.040	1.230
53	-.030	1.010	.050	1.010
54	.190	0.930	.270	0.910
55	.190	0.930	.270	0.910
56	.020	0.980	-.010	1.060
57	-.040	0.970	-.580	1.010
58	-.080	1.010	.000	1.000
59	-.090	1.010	-.010	1.000
60	.080	0.530	.160	0.520
61	.000	1.060	.080	1.050
62	-.020	1.050	.060	1.050
63	.020	0.960	-.380	1.020
64	-.040	0.890	-.150	0.980
65	.100	1.000	-.210	1.040
66	.040	1.150	.120	1.120
67	.050	0.920	-.480	1.030
68	-.070	1.010	-.340	1.020
69	.020	1.130	-.190	1.060
70	-.070	0.950	-.120	0.990
71	-.350	1.030	-1.100	1.010
72	-.110	1.000	-.030	0.990
73	.020	1.030	.100	1.020
74	.010	1.060	.090	1.050
75	.170	0.930	.050	0.990
76	.010	1.140	.100	1.110
77	-.010	1.020	.070	1.010
78	.100	1.370	.220	1.750
79	.050	1.040	.130	0.960

	Dichotomous		Partial Credit	
Item	Rasch Difficulty	Discrimination[a]	Rasch Difficulty	Discrimination
80	-.020	0.990	-.090	1.000
81	.030	1.170	.110	1.160
82	-.030	1.060	.050	1.050
83	-.020	1.050	.060	1.040
84	.010	1.120	-.330	1.080
85	-.050	0.970	-.300	1.000
86	.040	1.410	-.070	1.060
87	-.060	1.050	.020	1.050
88	.040	0.770	.120	0.730
89	.060	0.990	.140	0.950
90	.010	1.120	.100	1.110
91	-.020	1.020	.060	1.010
92	-.010	0.980	.070	0.950
93	-.020	0.890	.070	0.870
94	-.030	0.910	.050	0.890
95	.050	1.270	.130	1.210
96	-.080	1.030	.000	1.020
97	-.040	1.000	.040	0.980
98	-.130	1.000	-.470	1.010
99	.020	0.960	.100	0.940
100	.050	0.980	.130	0.910
101	.090	0.910	.170	0.900
102	-.030	1.020	.050	1.020
103	.060	0.460	.140	0.490
104	-.120	1.020	-.350	1.010
105	-.010	1.020	.070	1.000
106	-.070	1.020	.010	1.010
107	.000	1.020	.080	1.010
108	-.260	1.000	-.180	1.000
109	-.070	1.060	.010	1.060
110	-.120	1.000	-.040	1.000
111	.000	0.980	.010	1.030
112	.100	1.240	.180	1.110
113	-.140	1.010	-.060	1.010
114	-.050	1.030	.030	1.020
115	.060	1.220	.090	1.450
116	-.050	1.050	.040	1.050
117	.040	0.910	.120	0.860
118	.020	0.970	.100	0.960
119	.160	0.930	-.090	1.000

	Dichotomous		Partial Credit	
Item	Rasch Difficulty	Discrimination[a]	Rasch Difficulty	Discrimination
120	.000	1.020	.080	1.000
121	.070	0.750	.160	0.680
122	-.050	1.030	.030	1.030
123	.040	0.830	-.180	1.000
124	-.100	1.040	-.020	1.040
125	-.130	1.010	-.050	1.010
126	.190	0.840	.270	0.810
127	.150	0.940	-.020	1.030
128	-.010	0.970	-.250	1.010
129	-.170	1.000	-.090	1.000
130	.060	0.740	.140	0.670
131	-.070	1.010	.010	1.010
132	-.120	1.010	-.040	1.010
133	.060	0.480	.140	0.460
134	-.140	0.990	-.060	0.990
135	.090	0.840	.180	0.810
136	-.130	1.010	-.260	1.010
137	-.060	0.960	.020	0.950
138	.090	0.450	.170	0.390
139	.070	0.780	.160	0.690
140	-.020	1.030	-.150	1.040
141	.380	1.000	.340	1.000
142	.050	1.380	-.050	1.080
143	.070	0.680	.150	0.630
144	.030	0.890	.110	0.860
145	.050	0.890	.130	0.820
146	-.100	0.960	-.020	0.940
147	.030	0.530	-.190	0.910
148	.060	0.880	.140	0.850
149	.130	1.190	.030	1.080
150	-.020	1.080	.070	1.070
151	.040	0.890	.120	0.840
152	.170	0.830	.250	0.790
153	.090	0.450	.170	0.400
154	-.020	1.050	.000	1.030
155	.070	1.200	.150	1.170
156	-.120	0.990	-.230	1.000
157	-.100	1.000	-.020	0.990
158	-.170	0.990	-.090	0.990
159	.010	0.940	-.170	1.000

	Dichotomous		Partial Credit	
Item	Rasch Difficulty	Discrimination[a]	Rasch Difficulty	Discrimination
160	.080	1.320	.160	1.180
161	.110	1.130	-.320	1.060
162	.060	0.550	-.070	0.980
163	.130	0.790	-.060	0.980
164	.110	1.090	.190	1.030
165	.050	0.770	-.120	0.990
166	.140	0.730	.020	0.950
167	.140	0.960	.220	0.920
168	-.020	1.070	.060	1.070
169	-.150	0.980	-.070	0.980
170	-.110	0.980	-.030	0.970
171	.040	1.170	.120	1.110
172	-.070	1.020	.010	1.020
173	-.050	1.050	-.230	1.030
174	-.070	0.970	.010	0.960
175	.010	1.170	.090	1.180
176	-.010	1.060	.070	1.050
177	.080	1.150	-.070	1.060
178	-.070	1.000	.010	0.990
179	-.030	0.960	-.440	1.000
180	-.060	0.980	.020	0.970
181	-.070	1.050	.010	1.050
182	-.090	1.050	-.010	1.050
183	-.030	1.070	.050	1.070
184	-.100	0.970	-.020	0.960
185	.060	1.130	-.190	1.050
186	-.130	1.000	-.050	0.990
187	-.020	1.020	.060	1.020
188	.100	1.510	-.300	1.110

Note. [a]Winsteps estimates discrimination after calculating the Rasch difficulty

Appendix G: SME Ratings for Exam 1

	Q1				Q2				Q3				Q4				Grand
Item	R1[a]	R2	R3	Mean	R1	R2	R3	Mean	R1	R2	R3	Mean	R1	R2	R3	Mean	Mean
1	1.00	1.00	.75	.92	.50	1.00	.50	.67	.50	1.00	.75	.75	.67	1.00	.33	.67	.75
2	1.00	1.00	.75	.92	.75	1.00	.50	.75	1.00	1.00	.50	.83	.67	1.00	.67	.78	.82
3	1.00	1.00	.75	.92	1.00	1.00	.50	.83	1.00	1.00	.75	.92	1.00	1.00	.67	.89	.89
4	.75	1.00	.75	.83	.50	1.00	.50	.67	.75	1.00	.75	.83	.33	1.00	1.00	.78	.78
5	1.00	1.00	.75	.92	.50	1.00	.50	.67	1.00	1.00	.75	.92	.67	1.00	.67	.78	.82
6	1.00	1.00	.50	.83	.75	1.00	.25	.67	1.00	1.00	.50	.83	.67	1.00	.33	.67	.75
7	1.00	1.00	1.00	1.00	.50	1.00	.50	.67	1.00	1.00	.75	.92	.67	1.00	1.00	.89	.87
8	1.00	1.00	.75	.92	.75	1.00	.50	.75	1.00	1.00	.75	.92	.67	1.00	1.00	.89	.87
9	1.00	1.00	.75	.92	.50	1.00	.25	.58	.75	1.00	.75	.83	1.00	1.00	1.00	1.00	.83
10	1.00	1.00	.50	.83	.50	.50	.50	.50	1.00	1.00	1.00	1.00	.67	.67	1.00	.78	.78
11	1.00	1.00	.50	.83	.50	1.00	.50	.67	1.00	1.00	.75	.92	.67	1.00	.67	.78	.80
12	1.00	1.00	1.00	1.00	.50	1.00	.50	.67	.75	1.00	.75	.83	.67	1.00	1.00	.89	.85
13	1.00	1.00	.50	.83	.75	1.00	.50	.75	1.00	1.00	.75	.92	1.00	1.00	.67	.89	.85
14	1.00	1.00	1.00	1.00	.50	1.00	.50	.67	.75	1.00	1.00	.92	.67	1.00	.67	.78	.84
15	1.00	1.00	.75	.92	.50	1.00	.25	.58	.75	1.00	.75	.83	.67	1.00	.33	.67	.75
16	1.00	1.00	1.00	1.00	.50	1.00	.75	.75	1.00	1.00	1.00	1.00	1.00	1.00	.67	.89	.91
17	1.00	1.00	.75	.92	.50	1.00	.50	.67	.75	1.00	.75	.83	.67	1.00	.67	.78	.80
18	1.00	1.00	.75	.92	1.00	1.00	.50	.83	1.00	1.00	.50	.83	1.00	1.00	.67	.89	.87
19	.75	1.00	1.00	.92	.50	1.00	.25	.58	.75	1.00	1.00	.92	.33	1.00	.67	.67	.77
20	1.00	1.00	1.00	1.00	.50	1.00	.50	.67	.75	1.00	1.00	.92	.67	1.00	.67	.78	.84
21	1.00	1.00	1.00	1.00	.75	1.00	.50	.75	1.00	1.00	.75	.92	.67	1.00	1.00	.89	.89
22	1.00	1.00	.75	.92	.50	1.00	.50	.67	1.00	1.00	.75	.92	1.00	1.00	.67	.89	.85

	Q1				Q2				Q3				Q4				Grand
Item	R1[a]	R2	R3	Mean	R1	R2	R3	Mean	R1	R2	R3	Mean	R1	R2	R3	Mean	Mean
23	1.00	1.00	.75	.92	.75	1.00	.75	.83	1.00	.75	.75	.83	1.00	.67	.67	.78	.84
24	1.00	1.00	.75	.92	.75	1.00	.50	.75	1.00	1.00	.50	.83	1.00	1.00	.67	.89	.85
25	.75	1.00	.75	.83	.50	1.00	.50	.67	.50	1.00	.75	.75	.33	1.00	.67	.67	.73
26	1.00	1.00	.25	.75	.50	1.00	.25	.58	1.00	1.00	.50	.83	.67	1.00	.33	.67	.71
27	1.00	1.00	1.00	1.00	.50	1.00	.75	.75	1.00	1.00	.75	.92	.67	1.00	1.00	.89	.89
28	1.00	1.00	.75	.92	.50	1.00	.50	.67	1.00	1.00	.75	.92	.67	1.00	1.00	.89	.85
29	1.00	1.00	1.00	1.00	.75	1.00	.75	.83	1.00	1.00	.50	.83	1.00	1.00	1.00	1.00	.92
30	.75	1.00	.75	.83	.50	1.00	.50	.67	1.00	1.00	.75	.92	.67	1.00	.67	.78	.80
31	1.00	1.00	1.00	1.00	.75	1.00	.50	.75	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.94
32	1.00	1.00	.25	.75	.75	1.00	.25	.67	.75	1.00	.50	.75	.67	1.00	.33	.67	.71
33	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.75	.92	1.00	1.00	1.00	1.00	.98
34	1.00	1.00	1.00	1.00	.75	1.00	.75	.83	1.00	1.00	.75	.92	.67	1.00	1.00	.89	.91
35	1.00	1.00	1.00	1.00	.75	1.00	.75	.83	.75	1.00	.75	.83	.67	1.00	1.00	.89	.89
36	1.00	1.00	1.00	1.00	1.00	1.00	.75	.92	1.00	1.00	.75	.92	1.00	1.00	1.00	1.00	.96
37	1.00	1.00	.50	.83	.75	1.00	.25	.67	1.00	1.00	.50	.83	1.00	1.00	.67	.89	.81
38	1.00	1.00	1.00	1.00	.75	1.00	.75	.83	1.00	1.00	.75	.92	1.00	1.00	1.00	1.00	.94
39	1.00	1.00	.75	.92	.75	1.00	.50	.75	1.00	1.00	.75	.92	.67	1.00	.67	.78	.84
40	1.00	1.00	1.00	1.00	1.00	.50	1.00	.83	1.00	.75	1.00	.92	1.00	.67	1.00	.89	.91
41	1.00	1.00	1.00	1.00	.75	1.00	.75	.83	1.00	1.00	.75	.92	1.00	1.00	1.00	1.00	.94
42	1.00	1.00	1.00	1.00	.75	1.00	.75	.83	1.00	1.00	.75	.92	1.00	1.00	.67	.89	.91
43	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.33	.78	.94
44	1.00	1.00	1.00	1.00	.75	1.00	.75	.83	1.00	1.00	.75	.92	1.00	1.00	1.00	1.00	.94
45	1.00	1.00	.75	.92	.75	1.00	.50	.75	1.00	1.00	.75	.92	.67	1.00	.67	.78	.84
46	1.00	1.00	1.00	1.00	1.00	1.00	.75	.92	1.00	1.00	.75	.92	1.00	1.00	1.00	1.00	.96
47	1.00	1.00	.75	.92	.75	1.00	.50	.75	1.00	1.00	.75	.92	.67	1.00	1.00	.89	.87

	Q1				Q2				Q3				Q4				Grand
Item	R1[a]	R2	R3	Mean	R1	R2	R3	Mean	R1	R2	R3	Mean	R1	R2	R3	Mean	Mean
48	1.00	1.00	1.00	1.00	1.00	1.00	.75	.92	1.00	1.00	.75	.92	1.00	1.00	1.00	1.00	.96
49	.75	1.00	1.00	.92	.50	1.00	.50	.67	.50	1.00	.50	.67	.67	1.00	.67	.78	.76
50	1.00	1.00	1.00	1.00	1.00	1.00	.50	.83	1.00	1.00	.75	.92	1.00	1.00	1.00	1.00	.94
51	1.00	1.00	1.00	1.00	.75	1.00	.25	.67	.75	1.00	.75	.83	.67	1.00	.67	.78	.82
52	1.00	1.00	.75	.92	1.00	1.00	.25	.75	1.00	1.00	.75	.92	1.00	1.00	.67	.89	.87
53	1.00	1.00	.50	.83	.75	1.00	.50	.75	1.00	1.00	.50	.83	1.00	1.00	.67	.89	.83
54	1.00	1.00	1.00	1.00	1.00	1.00	.75	.92	1.00	1.00	1.00	1.00	1.00	1.00	.33	.78	.92
55	1.00	1.00	1.00	1.00	1.00	1.00	.75	.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.98
56	1.00	1.00	1.00	1.00	1.00	1.00	.75	.92	1.00	1.00	.75	.92	1.00	1.00	1.00	1.00	.96
57	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
58	1.00	1.00	1.00	1.00	.75	1.00	.50	.75	1.00	1.00	.75	.92	.67	1.00	1.00	.89	.89
59	1.00	1.00	.75	.92	.50	1.00	.50	.67	1.00	1.00	1.00	1.00	.67	1.00	.67	.78	.84
60	1.00	1.00	1.00	1.00	1.00	1.00	.75	.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.98
61	1.00	1.00	1.00	1.00	1.00	.75	.50	.75	1.00	.75	.75	.83	1.00	.67	1.00	.89	.87
62	1.00	1.00	.75	.92	.75	.75	.25	.58	1.00	.75	.75	.83	.67	.67	.67	.67	.75
Mean	.98	1.00	.84	.94	.73	.98	.56	.75	.93	.98	.76	.89	.81	.97	.78	.86	.86
SD	.07	.00	.20	.14	.19	.10	.20	.24	.14	.06	.15	.16	.20	.09	.22	.20	.20

[a]Note. R1, R2, and R3 designates Raters 1, 2 and 3 respectively.

Appendix H: SME Ratings for Exam 2

	Q1				Q2				Q3				Q4				Grand
Item	R1[a]	R2	R3	Mean	R1	R2	R3	Mean	R1	R2	R3	Mean	R1	R2	R3	Mean	Mean
1	1.00	.75	.25	.67	1.00	.50	.25	.58	1.00	.50	.50	.67	.67	.67	.67	.67	.65
2	1.00	.75	.25	.67	.25	.25	.00	.17	.75	.50	.50	.58	.33	.00	.00	.11	.38
3	1.00	.75	.25	.67	.75	.50	.25	.50	.75	.50	.50	.58	.67	.67	.33	.56	.58
4	1.00	.75	1.00	.92	.75	.75	1.00	.83	.75	.75	1.00	.83	.67	1.00	1.00	.89	.87
5	1.00	.50	.25	.58	.75	.50	.50	.58	.75	.50	.50	.58	1.00	.33	.33	.56	.58
6	1.00	.00	.50	.50	1.00	.50	.00	.50	.75	.00	.50	.42	1.00	.00	.67	.56	.49
7	1.00	.75	.25	.67	.75	1.00	.00	.58	.75	.50	.75	.67	.67	.33	.33	.44	.59
8	1.00	1.00	.00	.67	.75	.75	.25	.58	.75	.75	.50	.67	.67	1.00	.67	.78	.67
9	1.00	.00	.25	.42	.50	.50	.25	.42	.75	.50	.50	.58	1.00	.33	.67	.67	.52
10	1.00	.75	.25	.67	.75	.50	.00	.42	.75	.50	.50	.58	.67	.33	.67	.56	.56
11	1.00	.50	.50	.67	.75	.75	.50	.67	.50	.50	.25	.42	.67	.67	.33	.56	.58
12	1.00	.75	.25	.67	.75	.75	.25	.58	.50	.75	.25	.50	1.00	.67	.33	.67	.60
13	1.00	.75	.25	.67	.50	.75	.25	.50	.50	.75	.25	.50	1.00	.67	.33	.67	.58
14	1.00	.50	.00	.50	.75	.50	.25	.50	.50	.25	.25	.33	1.00	.33	.33	.56	.47
15	1.00	.75	.50	.75	.75	.50	.25	.50	.75	.75	.25	.58	1.00	.33	.33	.56	.60
16	1.00	.50	.50	.67	.50	.25	.25	.33	.50	.25	.25	.33	.67	.33	.33	.44	.44
17	1.00	.75	.50	.75	.50	.50	.25	.42	.75	.75	.25	.58	.67	.33	.33	.44	.55
18	1.00	.75	.50	.75	.50	.50	.25	.42	.75	.75	.25	.58	.67	.33	.33	.44	.55
19	1.00	.50	.50	.67	.50	.50	.00	.33	.75	.25	.25	.42	.67	.00	.33	.33	.44
20	1.00	.75	.25	.67	.50	.50	.25	.42	.75	.50	.25	.50	.33	.33	.33	.33	.48
21	1.00	.75	.25	.67	.50	.75	.00	.42	.75	.75	.25	.58	.67	1.00	.33	.67	.58
22	1.00	.75	.25	.67	.50	.75	.00	.42	.75	.75	.25	.58	.67	1.00	.33	.67	.58
23	1.00	.75	.25	.67	.75	1.00	.00	.58	1.00	.75	.25	.67	.67	1.00	.33	.67	.65

	Q1				Q2				Q3				Q4				Grand
Item	R1[a]	R2	R3	Mean	R1	R2	R3	Mean	R1	R2	R3	Mean	R1	R2	R3	Mean	Mean
24	1.00	.50	.50	.67	1.00	.50	.75	.75	.75	.50	.25	.50	.33	.33	.33	.33	.56
25	1.00	.50	.50	.67	1.00	.50	.50	.67	.75	.50	.25	.50	.67	1.00	1.00	.89	.68
26	1.00	.75	.75	.83	.50	.50	.25	.42	.75	.50	.25	.50	.67	.33	.33	.44	.55
27	1.00	.25	.25	.50	.75	.00	.25	.33	.75	.25	.25	.42	.33	.00	.33	.22	.37
28	1.00	.75	.50	.75	.50	.75	.25	.50	.75	.75	.25	.58	.33	1.00	.33	.56	.60
29	.75	.50	.25	.50	.75	.50	.25	.50	.75	.50	.25	.50	.33	.33	.33	.33	.46
30	1.00	.25	.50	.58	.75	.50	.25	.50	.75	.25	.75	.58	.67	.33	.67	.56	.56
31	1.00	.75	.50	.75	.50	.75	.25	.50	.75	.75	.75	.75	.67	1.00	.67	.78	.69
32	1.00	.75	.50	.75	1.00	.75	.25	.67	1.00	.75	.50	.75	.67	.33	.33	.44	.65
33	1.00	.50	.50	.67	.75	.50	.50	.58	.75	.25	.25	.42	.67	.33	.00	.33	.50
34	1.00	.00	.50	.50	.75	.50	.50	.58	.75	.50	.50	.58	1.00	.00	.33	.44	.53
35	1.00	.75	1.00	.92	1.00	.75	.75	.83	1.00	.75	.75	.83	1.00	1.00	1.00	1.00	.90
36	1.00	.50	.25	.58	.50	.50	.25	.42	.50	.25	.25	.33	.67	.33	.33	.44	.44
37	1.00	.50	.50	.67	.50	.25	.25	.33	.75	.50	.50	.58	.67	.33	.33	.44	.51
38	1.00	.75	1.00	.92	.50	.75	.75	.67	1.00	.75	.75	.83	.67	.67	1.00	.78	.80
39	.75	.50	.00	.42	.75	.50	.25	.50	.50	.25	.25	.33	.33	.33	.33	.33	.40
40	1.00	.75	.25	.67	.50	.50	.25	.42	.75	.50	.25	.50	.67	.33	.00	.33	.48
41	1.00	.50	.50	.67	.50	.50	.25	.42	.75	.25	.50	.50	.67	.33	.33	.44	.51
42	1.00	.00	.00	.33	.75	.50	.25	.50	.75	.00	.25	.33	.67	.33	.33	.44	.40
43	1.00	.50	.25	.58	.50	.50	.00	.33	.75	.25	.25	.42	.67	.33	.33	.44	.44
44	1.00	.75	.25	.67	.50	.50	.25	.42	.75	.50	.50	.58	.67	.33	.67	.56	.56
Mean	.99	.59	.39	.65	.66	.56	.28	.50	.74	.51	.40	.55	.68	.48	.43	.53	.56
SD	.05	.24	.24	.32	.18	.19	.22	.26	.13	.22	.19	.23	.20	.31	.24	.28	.28

[a]Note. R1, R2, and R3 designates Raters 1, 2 and 3 respectively.

Appendix I: SME Ratings for Exam 3

	Q1				Q2				Q3				Q4				Grand
Item	R1[a]	R2	R3	Mean	R1	R2	R3	Mean	R1	R2	R3	Mean	R1	R2	R3	Mean	Mean
1	1.00	1.00	.75	.92	1.00	1.00	.50	.83	1.00	1.00	1.00	1.00	1.00	1.00	.33	.78	.88
2	.75	1.00	1.00	.92	.50	1.00	.50	.67	.75	1.00	1.00	.92	.33	1.00	.67	.67	.79
3	.25	1.00	1.00	.75	.25	.50	.50	.42	.25	.50	1.00	.58	.33	1.00	.67	.67	.60
4	.25	1.00	1.00	.75	.50	.50	.75	.58	.25	.75	1.00	.67	.33	.67	1.00	.67	.67
5	.75	1.00	1.00	.92	.25	.50	.50	.42	.50	1.00	1.00	.83	.33	1.00	.67	.67	.71
6	.25	1.00	.75	.67	.00	.75	.50	.42	.50	.75	1.00	.75	.00	.33	.33	.22	.51
7	.75	1.00	.75	.83	.25	.75	.50	.50	.50	.75	1.00	.75	.33	1.00	.33	.56	.66
8	.25	.25	1.00	.50	.00	.75	.50	.42	.00	.25	1.00	.42	.00	.67	.67	.44	.44
9	1.00	1.00	.75	.92	.50	.75	.50	.58	.75	1.00	1.00	.92	1.00	1.00	.33	.78	.80
10	.75	1.00	1.00	.92	.75	1.00	1.00	.92	.75	1.00	1.00	.92	.33	1.00	1.00	.78	.88
11	1.00	1.00	1.00	1.00	.75	1.00	.50	.75	1.00	1.00	1.00	1.00	.67	1.00	1.00	.89	.91
12	.75	1.00	1.00	.92	.75	1.00	1.00	.92	.75	1.00	1.00	.92	1.00	1.00	1.00	1.00	.94
13	.50	.25	.75	.50	.25	.75	.50	.50	.50	1.00	1.00	.83	.33	.67	.33	.44	.57
14	1.00	1.00	.75	.92	.75	1.00	.25	.67	1.00	1.00	1.00	1.00	1.00	1.00	.67	.89	.87
15	.75	.75	.75	.75	.00	.50	.50	.33	.50	.00	1.00	.50	.33	.00	.67	.33	.48
16	.50	1.00	.75	.75	.50	.50	.50	.50	.50	.50	1.00	.67	.33	.67	.33	.44	.59
17	.00	1.00	.75	.58	.00	1.00	.25	.42	.00	.75	1.00	.58	.00	1.00	.67	.56	.53
18	.75	.75	.25	.58	.50	.75	.25	.50	.50	1.00	1.00	.83	.33	.67	.33	.44	.59
19	.25	1.00	.75	.67	.00	.50	.75	.42	.00	1.00	1.00	.67	.33	.67	.33	.44	.55
20	.00	.50	1.00	.50	.25	.75	.75	.58	.25	.50	1.00	.58	.33	.33	.67	.44	.53
21	.50	1.00	.75	.75	.25	.50	.50	.42	.50	.75	1.00	.75	.33	.67	.67	.56	.62
22	1.00	1.00	.50	.83	1.00	1.00	.50	.83	1.00	1.00	1.00	1.00	1.00	1.00	.33	.78	.86
23	1.00	1.00	1.00	1.00	1.00	1.00	.50	.83	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.96

	Q1				Q2				Q3				Q4				Grand
Item	R1[a]	R2	R3	Mean	R1	R2	R3	Mean	R1	R2	R3	Mean	R1	R2	R3	Mean	Mean
24	1.00	1.00	.75	.92	1.00	1.00	.50	.83	.50	1.00	1.00	.83	1.00	1.00	.67	.89	.87
25	1.00	1.00	.75	.92	1.00	1.00	.50	.83	1.00	1.00	1.00	1.00	1.00	1.00	.33	.78	.88
26	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
27	.75	1.00	.75	.83	.25	.50	.50	.42	.50	.50	1.00	.67	.33	.67	.33	.44	.59
28	.75	1.00	.50	.75	.00	1.00	.25	.42	.50	1.00	1.00	.83	.00	1.00	.33	.44	.61
29	.50	1.00	.25	.58	.50	.50	.25	.42	.50	.50	1.00	.67	.33	1.00	.33	.56	.56
30	.75	1.00	1.00	.92	.75	1.00	.50	.75	.75	1.00	1.00	.92	.33	1.00	1.00	.78	.84
31	1.00	1.00	1.00	1.00	.75	.75	.50	.67	.75	1.00	1.00	.92	1.00	1.00	.67	.89	.87
32	1.00	1.00	.75	.92	.75	.75	.50	.67	.75	.75	1.00	.83	.67	1.00	1.00	.89	.83
33	.75	1.00	1.00	.92	.75	.75	1.00	.83	.75	.75	1.00	.83	.67	1.00	1.00	.89	.87
34	.75	.75	1.00	.83	.50	.50	1.00	.67	.50	.50	1.00	.67	.67	.67	1.00	.78	.74
35	.50	1.00	1.00	.83	.50	1.00	.75	.75	.50	1.00	1.00	.83	.67	1.00	.67	.78	.80
36	.50	1.00	1.00	.83	.50	1.00	.75	.75	.50	1.00	1.00	.83	.67	1.00	.67	.78	.80
37	.75	1.00	.75	.83	.50	.50	.50	.50	.50	.75	1.00	.75	.67	.67	.33	.56	.66
38	.75	1.00	.75	.83	.50	1.00	.25	.58	.75	1.00	1.00	.92	.33	1.00	.33	.56	.72
39	.50	1.00	1.00	.83	.50	1.00	1.00	.83	.50	1.00	1.00	.83	.33	1.00	1.00	.78	.82
40	.75	1.00	1.00	.92	.75	1.00	1.00	.92	.75	1.00	1.00	.92	.67	1.00	1.00	.89	.91
41	.75	1.00	1.00	.92	1.00	.50	.75	.75	.75	.75	1.00	.83	.67	.67	.67	.67	.79
42	.75	1.00	1.00	.92	.50	.25	.50	.42	.75	.75	1.00	.83	.33	.67	.67	.56	.68
43	.50	1.00	.75	.75	.00	.25	.50	.25	.50	.25	1.00	.58	.00	.67	.33	.33	.48
44	.50	1.00	.75	.75	.50	.50	.50	.50	.50	.75	1.00	.75	.33	.67	.33	.44	.61
45	.75	1.00	.75	.83	.25	1.00	.50	.58	.75	1.00	1.00	.92	.33	1.00	.33	.56	.72
46	.50	1.00	.75	.75	.50	1.00	.25	.58	.50	1.00	1.00	.83	.00	1.00	.33	.44	.65
47	.50	1.00	.75	.75	.00	.50	1.00	.50	.50	.50	1.00	.67	.00	.67	.33	.33	.56
48	.75	1.00	.75	.83	.75	.50	.50	.58	.75	.75	1.00	.83	.67	.67	.33	.56	.70

	Q1				Q2				Q3				Q4				Grand
Item	R1[a]	R2	R3	Mean	R1	R2	R3	Mean	R1	R2	R3	Mean	R1	R2	R3	Mean	Mean
49	.50	1.00	.75	.75	.50	.75	.50	.58	.50	.75	1.00	.75	.67	1.00	.33	.67	.69
50	.75	1.00	1.00	.92	.75	1.00	.25	.67	.75	1.00	1.00	.92	.67	1.00	.67	.78	.82
51	.75	1.00	1.00	.92	.75	1.00	1.00	.92	.75	.75	1.00	.83	.67	1.00	1.00	.89	.89
52	.75	1.00	1.00	.92	.75	1.00	.75	.83	.75	.75	1.00	.83	.67	1.00	1.00	.89	.87
53	.75	1.00	1.00	.92	.50	.75	.50	.58	.50	.75	1.00	.75	.33	1.00	1.00	.78	.76
54	.50	1.00	1.00	.83	.25	.50	.50	.42	.50	.75	1.00	.75	.33	.67	.33	.44	.61
55	1.00	1.00	1.00	1.00	.75	1.00	1.00	.92	.75	.50	1.00	.75	.67	1.00	1.00	.89	.89
56	.75	1.00	1.00	.92	.75	.75	.50	.67	.75	.75	1.00	.83	1.00	1.00	1.00	1.00	.85
57	1.00	1.00	1.00	1.00	.75	1.00	1.00	.92	1.00	.75	1.00	.92	1.00	1.00	.67	.89	.93
58	.50	1.00	1.00	.83	.50	.50	.50	.50	.50	.25	1.00	.58	.33	.00	.67	.33	.56
59	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Mean	.68	.95	.86	.83	.53	.78	.60	.63	.62	.80	1.00	.81	.53	.85	.64	.67	.74
SD	.26	.16	.18	.23	.31	.23	.24	.28	.25	.24	.00	.25	.32	.24	.28	.31	.28