The Dissertation Committee for Mishan G. B. Jensen Certifies that this is the approved version of the following dissertation:

Extension of the Item Pocket Method Allowing for Response Review and Revision to a Computerized Adaptive Test using the Generalized Partial Credit Model

Committee:

Tiffany A. Whittaker, Supervisor

S. Natasha Beretvas

Barbara G. Dodd

Matthew A. Hersh

Keenan A. Pituch

Extension of the Item Pocket Method Allowing for Response Review and Revision to a Computerized Adaptive Test using the Generalized Partial Credit Model

by

Mishan G. B. Jensen

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August, 2017

Dedication

For my father and daughter, Samuel Behrend and Ava Jensen. My dedication to completion of this Doctoral degree has been driven by the desire to make my father proud, even though he is not here to see it, and to show my daughter that anything is possible if you put your mind to it.

Acknowledgements

My interest in psychometrics, specifically computerized adaptive testing, began with Dr. Dodd's courses. I am deeply grateful to Barbara Dodd for her guidance and encouragement over the course of my graduate studies and through the qualifying process, as well as her support and participation on the dissertation committee. She has always been extremely generous with her time and advice, for which I am profoundly grateful.

I would like to thank Tiffany Whittaker, my dissertation chair, who has been very generous with her time and agreeing to chair my dissertation which was out of her specialty. Throughout the course of my studies, she has been a great source of knowledge as well as an outstanding advisor and teacher. I would also like to acknowledge and thank the rest of the committee members, Tasha Beretvas, Keenan Pituch, and Matthew Hersh. Through their courses I gained a deep understanding of quantitative methods and applications. Their suggestions have aided me in this dissertation research and their advice has been invaluable throughout the process.

I would like to acknowledge the support and love I have received from my family. My mother has provided unending support throughout my graduate studies and provided an ear to let me talk things out even though she may not understand anything I was saying. I would like to thank my husband for providing support by allowing many hours for me to study and write, without his support this dissertation would not have been possible. I would also like to thank my daughter, Ava, for giving up time with me so that I could finish my studies. I know I have missed out on a lot of playing but I am ready to make up for the lost time!

Extension of the Item Pocket Method Allowing for Response Review and Revision to a Computerized Adaptive Test using the Generalized Partial Credit Model

Mishan G. B. Jensen, Ph.D.

The University of Texas at Austin, 2017

Supervisor: Tiffany A. Whittaker

Computerized Adaptive Testing (CAT) has increased in the last few decades, due in part to the increased use and availability of personal computers, but also partly due to the benefits of CATs. CATs provide increased measurement precision of ability estimates while decreasing the demand on examinees with shorter tests. This is accomplished by tailoring the test to each examinee and selecting items that are not too difficult or too easy based on the examinees' interim ability estimate and responses to previous items. These benefits come at the cost of the flexibility to move through the test as an examinee would with a Paper and Pencil (P & P) test. The algorithms used in CATs for item selection and ability estimation require restrictions to response review and revision; however, a large portion of examinees desire options for review and revision of responses (Vispoel, Clough, Bleiler, Hendrickson, and Ihrig, 2002). Previous research has examined response review and revision in CATs with limited review and revision options and are limited to after all items had been administered. The development of the Item Pocket (IP) method (Han, 2013) has allowed for response review and revision

during the test, relaxing the restrictions, while maintaining an acceptable level of measurement precision. This is achieved by creating an item pocket in which items are placed, which are excluded from use in the interim ability estimation and the item selection procedures. The initial simulation study was conducted by Han (2013) who investigated the use of the IP method using a dichotomously-scored fixed length test. The findings indicated that the IP method does not substantially decrease measurement precision and bias in the ability estimates were within acceptable ranges for operational tests.

This simulation study extended the IP method to a CAT using polytomously-scored items using the Generalized Partial Credit model with exposure control and content balancing. The IP method was implemented in tests with three IP sizes (2, 3, and 4), two termination criteria (fixed and variable), two test lengths (15 and 20), and two item completion conditions (forced to answer and ignored) for items remaining in the IP at the end of the test. Additionally, four traditional CAT conditions, without implementing the IP method, were included in the design. Results found that the longer, 20 item IP method conditions using the forced answer method had higher measurement precision, with higher mean correlations between known and estimated theta, lower mean bias and RMSE, and measurement precision increased as IP size increased. The two item completion conditions (forced to answer and ignored) resulted in similar measurement precision. The variable length IP conditions resulted in comparable measurement precision as the corresponding fixed length IP conditions. The implications of the findings and the limitations with suggestions for future research are also discussed.

Table of Contents

List of Tables

List of Figures

Chapter I: Introduction

In the past few decades, with the increase of personal computers in daily life, computerized adaptive testing (CAT) has become pervasive. CAT has allowed for more accurate measurement of ability with shorter more efficient tests by tailoring the test to individual examinees. To accomplish this examinee-tailored test, if the examinee answered the first administered question right, the next administered question would be a little harder. If instead the first administered question was answered incorrectly, the next administered question would be a little easier. The difficulty of each subsequent administered item is based on the responses to previous items. This produces shorter more efficient tests because examinees are not given items that are too easy or too hard for them to answer. For these reasons, CATs are appealing to educators. However, examinees are restricted from moving through a CAT as they would with a paper-and-pencil test (P & P) in which they could skip items they were unsure of or review and revise answers. This restriction is necessary in CAT due to the algorithms used to estimate examinee ability which requires ability re-estimation after each question is answered.

A large portion of tests that students are exposed to in everyday life are of the P & P variety. Throughout school, from first grade through college, students are taught to skip and mark questions they are unsure of to return to later, if time allows. Additionally, it is suggested that they go back over the entire test when completed in order to check for careless mistakes. This is not allowed with CATs that adapt at the item level, due to the ability estimation procedures used. It is possible that due to these restrictions, measurement error is increased. Examinees may randomly choose an answer in order to move forward in the test because they believe that it may take too long to answer it and they are unsure of how much time they need to

complete the test. This type of random guessing does not add anything but measurement error to the examinee's ability estimate. Additionally, careless mistakes, such as selecting A when B was meant or moving a decimal place, cannot be corrected. Again, these mistakes would increase the measurement error and begs the question of what the test is measuring: the intended knowledge of the academic content area or test taking ability. In spite of these concerns, limited research on CATs that allow for review and response revision have been conducted. A large reason for the lack of research in this area is the concern for increasing the opportunity to cheat. Wainer (1993) has suggested that allowing response revision could open the door for cheating if examinees are knowledgeable about the CAT algorithms used. Regardless of these concerns, some researchers see the possible benefits of review and revision as outweighing the possibility of cheating.

Early research on CATs that allowed review and revision of items utilized licensure and certification assessments where a minimum level of competency was needed, resulting in a pass/fail decision. Lunz, Bergstrom, and Wright (1992) investigated the effect of review and revision on the efficiency of the CAT and the resulting ability estimates of the examinees with a sample of college students randomly assigned to conditions. Lunz, et al. (1992) created a CAT item bank from P & P forms of a medical technology certification examination. The length of the CAT varied with completion of at least 50 items and a maximum of 240 items, covering six content areas according to the pre-existing P & P test specification. The test length varied, with at least 50 items completed and a maximum of 240 items. The test would stop after the examinee's ability was above or below the pass/fail point by more than 1.3 times the standard error of measurement. Two review conditions were examined wherein examinees were allowed to review their responses or were not allowed to review their responses. The distinction between the two conditions is that the review condition allowed for revision of answers after all of the

items had been answered. The results indicated that after review and revision, the decrease in the efficiency of the CAT was minimal (1 %), within a standard error of measurement. The decision accuracy, pass or fail, was also comparable to the no review condition.

Stone and Lunz (1994) extended the previous study by Lunz et al. (1992) with a more comprehensive examination of the impact of response review and revision on the psychometric properties of two different CAT licensure tests using two different examinee populations for each test in a live testing situation. The test stopped after 50 items if the examinee's ability estimate was outside of the 95% confidence interval around the pass/fail point. If the ability estimate was within the confidence interval after 50 items, another 50 items would be completed. After 100 items, if the examinee's ability estimate was still included within the confidence interval, the pass/fail decision would be based on the current location in reference to the pass/fail point. This study, similar to the previous study by Lunz et al. (1992), allowed review and revision only after all of the items had been completed. The results indicated that allowing review and response changes after the initial CAT was completed minimally biased ability estimates. Decision accuracy after review and revision for both tests was 94% and 95% for Test 1 and Test 2, respectively. That is, only 6% of the examinees taking Test 1 changed their pass/fail decision by revision of their answers and only 5% of the examinees taking Test 2 changed their pass/fail decision following a revision of their answers. These examinees who were able to change the pass/fail decision were within one standard error of the pass/fail cut point. Examinees this close to the cut point, confidence in the pass/fail decision is minimal, at best, whether review and revision is allowed or not.

Based on these optimistic results, Stocking (1997) expanded on Lunz, et al. (1992) and Stone and Lunz's (1994) line of research with a simulation study including a more thorough

investigation of restricted review and revision options, with conditions replicating the 1994 findings. Stocking (1997) developed and investigated three restricted review models with limited success. These models included blocks of items, or sets of items grouped together. The conditions varied the number of items contained in a block and the number of blocks within each test. The least restrictive condition contained more items per block, with fewer blocks, which resulted in biased ability estimates. In contrast, the more restrictive conditions, which contained fewer items per block with more blocks of items per test, resulted in minimally biased ability estimates.

Vispoel, Hendrickson, and Bleiler (2000) extended this line of research in a live testing situation to investigate examinees attitudes about opportunities to review and change responses. Results supported Stocking's (1997) findings of limited bias in ability estimates when review and response changes were limited to small blocks of items. Examinees attitudes indicated that the majority of examinees desired an opportunity to review and revise answers, regardless of whether they utilized the option (Vispoel, Hendrickson, & Bleiler, 2000).

Vispoel, Clough, Bleiler, Hendrickson, and Ihrig (2002) investigated in a live testing situation, whether examinees can positively bias their ability estimates in CATs that allow review and response revision. The authors were concerned that examines could bias their ability estimates if they understood the CAT algorithm used by evaluating the difficulties of two consecutive items. If an item was answered incorrectly, the next item would be easier and the answer to the previous item should be changed. Likewise, if the next item was more difficult, then it could be assumed that the previous item was answered correctly. Vispoel, et al. (2002) evaluated the examinees' ability to distinguish differences in difficulties between items in a live testing situation with two conditions, one where the strategy was taught and used and another

where no strategy was taught. The results indicated that examinees in the strategy condition did not improve their score compared to the examinees in the no strategy condition. Moreover, the examinees in the review condition actually reduced their test scores due to errors when determining the difficulty between two consecutive items.

Until 2013, the research addressing the restriction of response review and change was limited to the previously described variations of restricted review and response revisions. All of the variations restricted response review and revision to after the CAT was completed; however, the number of blocks of items and number of items per block allowed for review and revision was varied. Similarly, all of the previous studies administered items that were scored dichotomously (either correct or incorrect) and utilized the simplest Item Response Theory (IRT) model, the 1-Parameter Logistic Model (1-PL).

Han (2013) developed the item pocket (IP) method to address these restrictions in a more flexible manner than previous research. This new method provides a pocket for placing items to skip and return to later. Any item placed in this pocket is not used to estimate the interim ability level until the item is removed from the pocket by finalizing an answer. The IP method has been shown to provide more accurate estimates of ability with less bias as compared to Stocking's (1997) results. Currently, this is the only known method to allow for response review and revision during a CAT, rather than after all of the items have been answered. However, this method has only been examined under a few simulated conditions. Han (2013) applied the IP method to a dichotomous CAT, similar to the previous research. Operational tests are often comprised of both items that are scored dichotomously and items that have more than two score categories, or referred to as polytomous items. These items require a constructed response and can be scored in a partial credit fashion. Han's (2013) IP method resulted in ability estimates

within an acceptable range of accurate measurement when items are scored dichotomously, producing a more flexible procedure that allows for response review and revision. Before the extension to mixed format CATs (CATs with both dichotomous and polytomous items), the performance of the IP procedure under the polytomous case needs to be investigated.

Based on Han's (2013) optimistic results this dissertation research extends the IP method to a polytomous IRT model, the Generalized Partial Credit model (Muraki 1992), that is appropriate for partial scoring. The study investigated three IP size conditions: (1) two items, (2) three items, and (3) four items. The IP method was implemented on both a fixed length and a variable length test. Test length was varied with test length of 15 and 20. Content balancing allows for multiple content areas to be covered in one test and item exposure control ensures that not all examinees receive the same items. Content balancing was not utilized in Han's (2013) study but was implemented in all conditions in the current study, thereby more closely approximating operational tests. Additionally, two item completion conditions for items that are left in the pocket at the end of the test, (1) forced answer and (2) ignored, were included in the study. The current study is a fully crossed factorial design (3 x 2 x 2 x 2), resulting in 24 conditions with 1,000 simulees and 500 replications per condition. Additionally, four baseline conventional CATs without implementing the item pocket were used as a comparison in evaluating the performance of the item pocket method, resulting in a total of 28 conditions.

The simulees' responses were generated from a normal distribution using IRTGEN (Whittaker, Fitzpatrick, Williams, & Dodd 2003). The item pool used is based on a national testing program and contains 157 items from three content areas with possible score points of 2, 3, and 4. Maximum Likelihood Estimation (MLE) was used to estimate simulees' ability with the use of variable step size adjustment when the ability estimate cannot be estimated.

Evaluation of the IP method applied to a CAT using the GPCM compared all conditions to each other and to the a traditional baseline CAT, with exposure control and content balancing but without an item pocket. Four main research questions were addressed in this dissertation research: 1) What is the impact of the IP method on precision of measurement across the range of ability levels when applied to a CAT using the GPCM with content balancing and exposure control procedures? 2) What is the impact on precision of measurement under the two termination criteria (i.e., fixed and variable length)? 3) What is the impact of the two item completion conditions (forced answer or ignored) on precision of measurement? 4) What impact does implementing the IP method have on test efficiency in the variable length conditions? The results of the simulated CAT using the IP method were analyzed in terms of item pocket usage, the overall precision of measurement in the final theta estimates, and test efficiency. The findings of this study will illustrate the applicability of this new method to a broader variety of CATs.

The following sections summarize the psychometric theory behind CAT and assumptions, as well as detailed descriptions of common models used in research and practice. Next, the components of CATs and relevant research on the performance of procedures are reviewed. The last section summarizes the research to date on CATs that allow for response review and revision, as well as describes the current simulation study.

Chapter 2: Literature Review

The present study extends a new method for Computerized Adaptive Testing (CAT), which allows for item review and revision, to tests that contain items that have multiple possible score points. These types of items can receive partial credit scoring. It is important to understand the models used in CAT and the components of CATs that are commonly used to understand the need and benefits of this new method.

This chapter begins with an introduction and review of Item Response Theory (IRT) and models. These models are used in CAT and without them, adaptive testing would not be possible. Following the IRT review, the components of CAT are reviewed as well as various procedures and algorithms that are commonly used in CAT.

Item Response Theory

Item Response Theory (IRT) was developed as an objective way of measuring latent traits, such as depression or math ability that cannot be directly measured. In Classical Test Theory, the item statistics are dependent on the group that took those items and scores are dependent on the test taken. IRT, originally called Latent Trait Theory, disentangles person parameters from item or test parameters. This is achieved by meeting much stronger assumptions than required by Classical Test Theory.

IRT Assumptions

There are three basic assumptions common to many IRT models: unidimensionality, local independence, and the correct specified functional form. For most IRT models, it is assumed

that there is only one underlying latent trait being measured, one dimension or unidimensionality, for which one ability estimate (θ) will be calculated for each examinee.

The second assumption, local independence, follows if the dimensionality assumption is met. There are two forms of local independence, weak and strong. The weak form of local independence holds when the item responses for a particular theta (θ), or ability level, are uncorrelated. The strong form of local independence holds when the item responses, conditional on theta, are statistically independent, meaning that there is no relationship among items, linear or non-linear. This is achieved when the probability of a correct response on one item is not influenced by the probability of a correct response on another item while controlling for theta (Embertson & Reise, 2000).

The third assumption of IRT is that the functional form must be correctly specified. The functional form is the mathematical relationship between the probability of a correct response and theta, represented by an Item Characteristic Curve (ICC). These ICCs depict the change in the probability of a correct response as it relates to changes in ability level (θ) (Embertson & Reise, 2000). The shape of the ICCs, for dichotomous models, is a function of the item parameters, such as difficulty, discrimination, and pseudo-guessing in the specified model. The shape, location on the theta scale, and the lower asymptote of the ICCs depict the item parameters specified in the model.

Dichotomous IRT Models

There are numerous IRT models, which are classified by the way the item is scored. For dichotomous models, the predicted probability of a response, conditional on ability level (θ), is

based on two possible responses, correct (1) or incorrect (0). The three most commonly used dichotomous IRT models are described below.

One-Parameter Logistic Model

The first and simplest dichotomous IRT model is the 1-Parameter Logistic model (1PL), also referred to as the Rasch model (Rasch, 1960). It is called the 1PL model because only one item parameter is included, the difficulty (*b*) parameter. The probability (P_i) of a correct response ($u_i = 1$) to an item *i* by an examinee with a given theta (θ) is defined as:

$$P_i\left(u_i = 1 \middle| \theta\right) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}}, \quad (1)$$

where the natural antilog of the difference between the examinees' ability level (θ) and the item's difficulty (*b*) is divided by one plus the natural antilog of the difference between theta and the item's difficulty. This equation represents the odds of success given the examinee's ability level and the item's difficulty divided by one plus the odds of success. When this probability is plotted against ability level, it generally produces an S-shaped logistic curve. In the 1PL model, items only differ in terms of difficulty, so all ICCs should have the same slope and a lower asymptote of 0, but differ in location on the difficulty scale. Figure 1 displays ICCs for 3 items that differ in difficulty only.

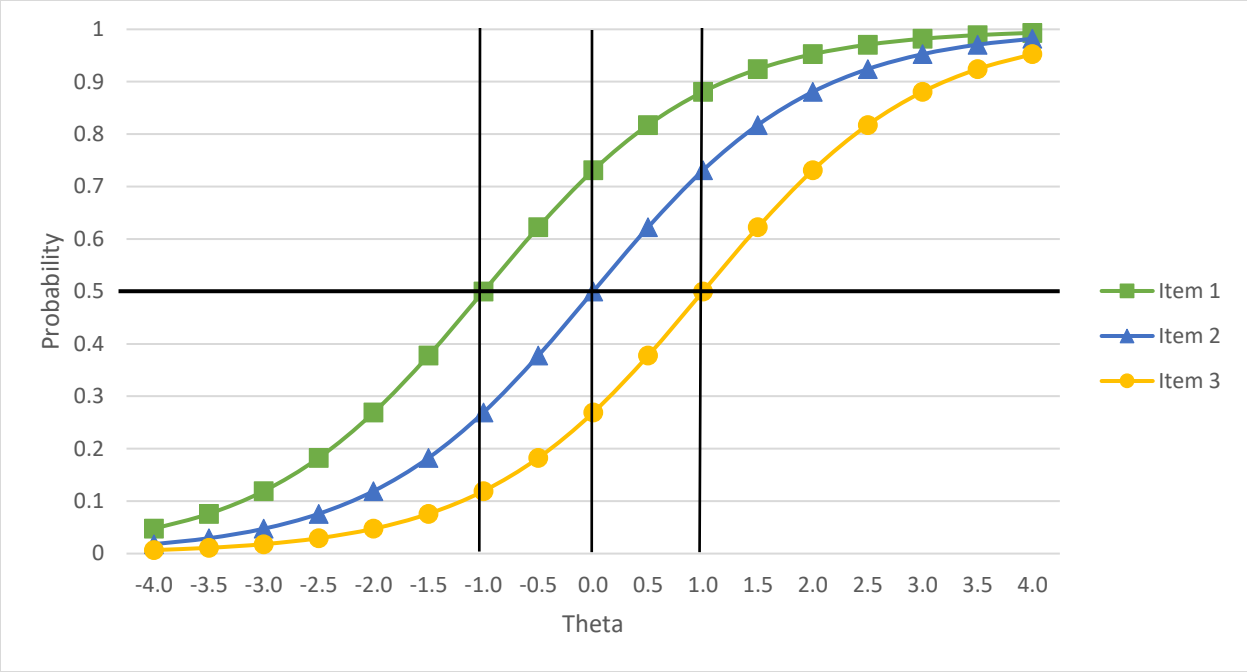

Figure 1. Item Characteristic Curves for a 1PL model.

The difficulty parameter (*b*) is depicted by the location of the point of inflection on the theta scale. The point of inflection is the point on the ICC where the rate of change is the highest, which corresponds to a 0.5 probability of a correct response. The line on the figure above at 0.5 probability of a correct response corresponds to the point of inflection for the three items' ICCs. Following the line down from the point of inflection corresponds to each item's difficulty (*b*) for items 1, 2, and 3 of $b = -1$, $b = 0$, and $b = 1$, respectively. The difficulty parameter is on the same scale as theta, which typically ranges from -4 to +4 with negative values indicating easier items and positive values indicating more difficult items.

Two-Parameter Logistic Model

The 2-Parameter Logistic model (2PL; Birnbaum, 1958), as the name implies, includes two item parameters: difficulty (*b*) and discrimination (*a*). Again, the probability (P_i) of a correct response ($u_i=1$) to an item *i* by an examinee with a given theta (θ) is defined as:

$$P_i\left(u_i = 1 \middle| \theta\right) = \frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}} \quad , \tag{2}$$

where the discrimination parameter (a) is proportional to the slope at the point of inflection where the probability of a correct response is 0.5. The a parameter is a multiplier to the difference between the current theta level for an examinee and the difficulty of the item. The multiplicative effect of the item's discrimination on the difference between the ability level and item difficulty has a stronger impact on the probability of a correct response when discrimination is high (Embertson & Reise, 2000). When the ICCs for a 2PL model are plotted, now the curves differ in terms of location on the ability scale and slope of the curves; however, the lower asymptotes should still originate at 0. Figure 2 displays ICCs for a 2PL model, where items differ in difficulty and discrimination.

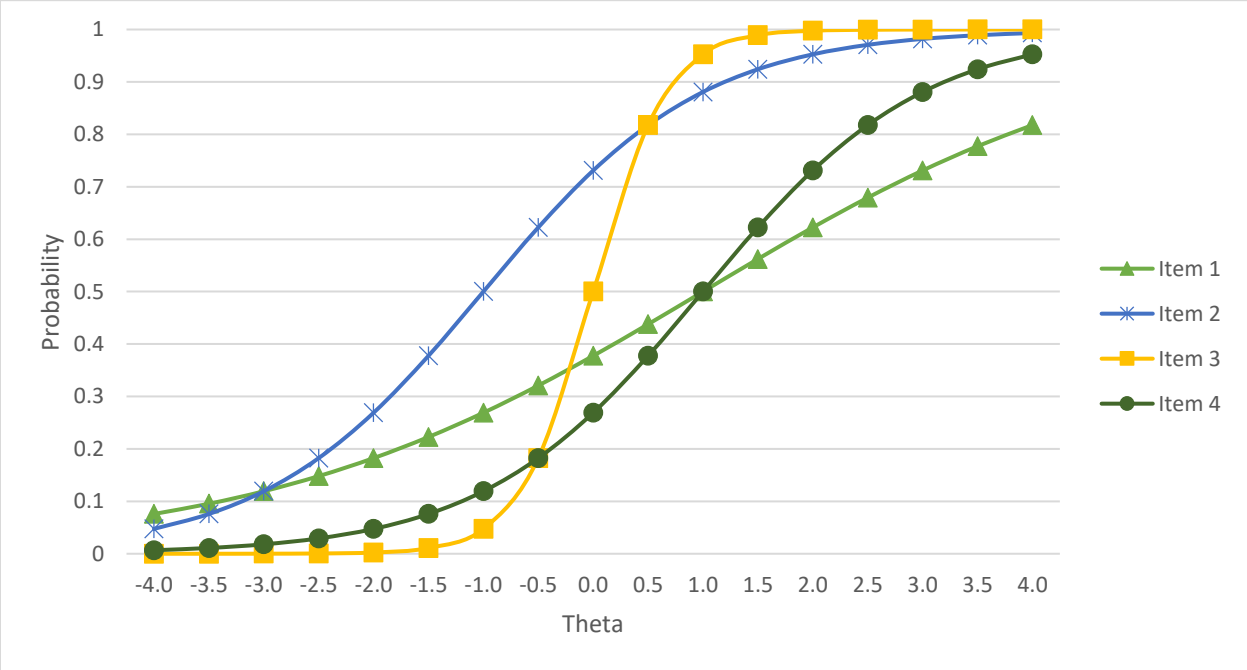

Figure 2. Item Characteristic Curves for 2PL model.

Generally, dichotomously (0/1) scored items will have an S-shaped ICC, however, the slope of the S depicts the discrimination parameter (*a*) or the rate of change in the probability of success at the point of inflection for a given ability level. The multiplicative effect of the difference between examinee's ability level and item difficulty can be seen, with the more discriminating items having steeper slopes at the point of inflection, indicating a greater impact on the probability of success for these items. Item 1 in Figure 2 has the lowest discrimination with a = 0.5, as can be seen with the most gradual slope at the point of inflection. Items 2 and 4 have identical slopes (a = 1) and item 3 has the highest discrimination value (a = 3), which corresponds to the steepest slope of the four items in Figure 2.

Three-Parameter Logistic Model

The 3-Parameter Logistic model (3PL; Birnbaum, 1968) extends the 2PL by adding a pseudo-chance parameter (c) to account for possible correct guessing on an item. However, the value for this parameter is generally less than what would be expected by random guessing. The probability (P_i) of a correct response (u_i=1) to an item i, conditional on ability level (θ), is defined as:

$$P_i\left(u_i = 1 \middle| \theta\right) = c_i + \left(1 - c_i\right)\frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}}\ , \tag{3}$$

where c_i is the pseudo-guessing parameter and is defined as the lower asymptote. The point of inflection is no longer at 0.5 probability. It is found by $(1 + c_i)/2$, which adjusts the point of inflection to account for the increase in the lower asymptote. The ICCs for a 3PL model can

now differ in location on the ability scale, slope of the curve, and lower asymptote. Figure 3 displays ICCs for items calibrated with a 3PL model.

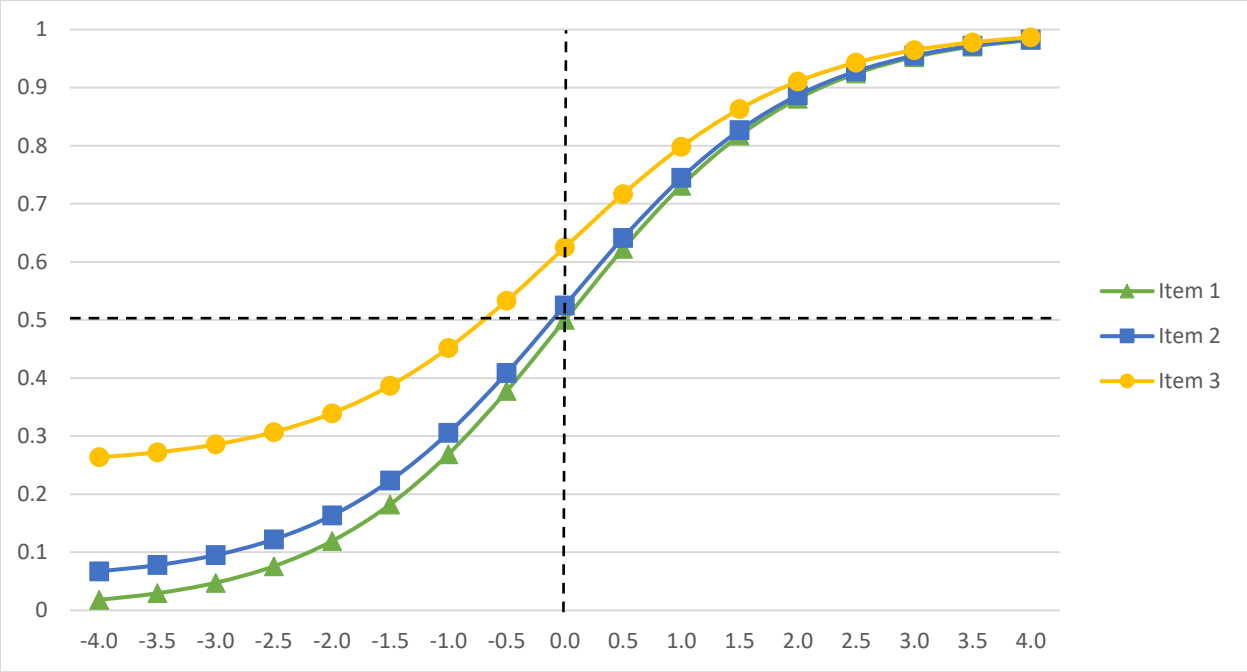

Figure 3. Item Characteristic Curves for 3PL model.

The lower asymptote of the ICC is the intersection point of the curve with the Y-axis, which impacts the point of inflection. When the intersection point is zero, as is seen above for the item marked with a triangle, the point of inflection is unchanged, represented by the dashed line in Figure 3. When the point of inflection is zero, there is a 0.5 probability of a correct response and the pseudo-guessing (c) parameter is equal to 0. When the lower asymptote is above zero, the c parameter is no longer zero, and the point of inflection is shifted up. This is represented in Figure 3 for the items represented by the square and the circle, which have c parameter values of 0.05 and 0.25, respectively. This change in the lower asymptote indicates that at the lower ability levels (θ), the probability of a correct response is above zero due to guessing. This more

complex model includes a greater number of parameters, with the possible consequence of non-convergence due to the estimation of the pseudo-guessing parameter (Embretson & Reise, 2000).

Polytomous IRT Models

Items that have more than two categories are appropriate for polytomous IRT models. The last 30 years have seen the development of many new polytomous IRT models. This family of models is composed of three major types of models classified by the procedure for calculating the probability of a response in a particular category. Dodd, de Ayala, and Koch (1995) surveyed the most common models used in CAT research, specifically the Difference Models and the Divide-By-Total Models. The third type of model, left-side added divide-by-total, is a nominal class of models that provides an undecided category for estimating a parameter for truly undecided participants (Dodd, de Ayala, & Koch, 1995). The third type of model has not been used in CAT research, so it is not discussed below. Three commonly used models that are appropriate for partial credit scoring are discussed in detail below. The summary of some of the more common models is not intended to be exhaustive. As such, readers are referred to additional references for information about models not discussed due to their inapplicability to the proposed research.

Difference Models

Difference models, as the name alludes, calculates probabilities for responses in a particular category by subtracting the probability of responding in adjacent categories, conditional on theta. These models require a two-step process for the calculation of probabilities: the first of which calculates the probability of a response in category *x* or higher for each category, P* functions, and the second step is the subtraction of adjacent category's P*

functions, conditional on theta. A common difference model appropriate for partial credit scoring is the Graded Response Model (Samejima, 1969). This model is appropriate for items that have multiple categories that are ordered in terms of correct steps to a solution (Dodd, de Ayala, & Koch, 1995). Another common difference model is Muraki's (1990) Rating Scale Model, however, the current study's focus is partial credit scoring of constructed responses to math problems, where application of the Rating Scale Model would not be appropriate and therefore not discussed further (see Muraki, 1990).

Graded Response Model

Samejima's (1969) Graded Response Model (GRM) was envisioned to handle partial credit scoring of items, where the higher categories indicate more correct steps toward a solution. The item's categories are ordered into $m_i + 1$ categories, where each category score for item i is a successive integer. The first stage, which calculates the probability of a particular category score or above, conditioned on theta (θ), for each possible category score, the P* functions, are defined as:

$$P_{ix}^{*}(\theta) = \frac{\exp\left[a_i(\theta - b_{ix})\right]}{1 + \exp\left[a_i(\theta - b_{ix})\right]}, \tag{4}$$

where a_i is the discrimination power for item i and b_{ix} is the boundary for a particular category score for item i. Each item includes a discrimination parameter, as well as, m_i category boundaries between $m_i + 1$ categories. The second stage takes the difference between adjacent P* functions to obtain the probability of a response in a particular category, found with this equation:

$$P_{ix}(\theta) = P^{*}_{ix}(\theta) - P^{*}_{ix+1}(\theta), \tag{5}$$

where the lowest possible category score function (P_0^*) is equal to 1.0 and the highest category score (P^*_{x+1}) is equal to 0. Plotting the P^* functions produces operating characteristic curves for each category, which determines the location of each of the category boundaries, depicted in Figure 4.

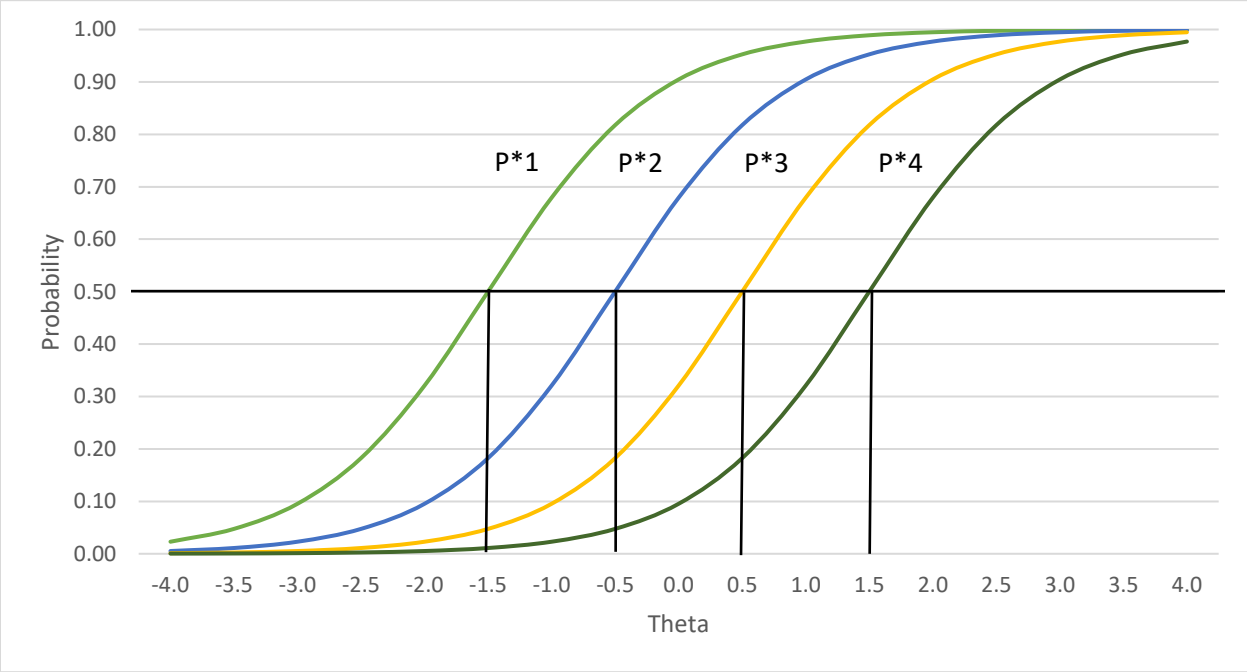

Figure 4. Operating Characteristic Curve for a 5 Category Item Under GRM.

Each of the P^* functions determine the location on the ability level scale at the point of inflection where the examinee has a 0.5 probability of responding above the threshold, or category boundary. Plotting the probability for each possible category score against theta produces Operating Characteristic Curves (OCC), with the point of inflection located at the point where there is a 0.5 probability of responding in a given category. Figure 5 displays a category response curve (CRC) for a five category item calibrated with the GRM.

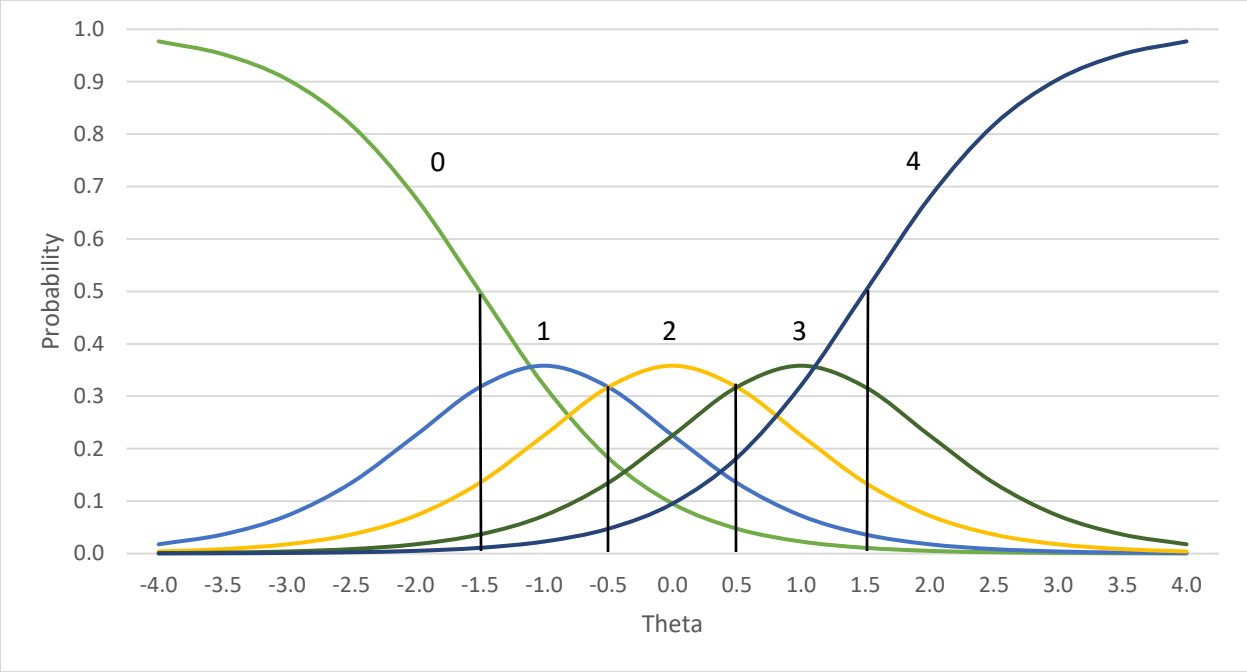

Figure 5. Category Response Curve using the GRM.

In the CRC shown above, the category thresholds obtained from Figure 4 are indicated by the black lines. This represents the location on the ability level scale where the examinee has a 0.5 probability of responding in the adjacent category. When items have only two categories, the GRM simplifies to the 2PL model.

Divide-by-Total Models

Contrary to difference models that calculate probabilities indirectly through the two step process, divide-by-total models find probabilities directly. As one would assume from the name, divide-by-total models find the probability of a particular category response for an item, conditioned on the examinee's theta, by dividing the exponential of the response category of interest by the sum of all categories' exponentials for the item. Another contrast to the difference models, the divide-by-total models do not require the difficulty, *b*-parameters, of each step to the solution to be ordered, meaning step 2 could be easier than step 1. The most general

of the divide-by-total models is the Nominal Response Model (Bock, 1972). Bock (1972) developed this model for use with multiple choice items where distractors are not easily ordered in terms of correctness, which is not an appropriate model to use for partial credit scoring (see Bock, 1972). The Successive Intervals Model (Rost, 1988) and Andrich's Rating Scale Model (Andrich, 1978) were developed for use with Likert measures, where responses are on an ordered continuum indicating the degree of agreement with a statement. Again, these models are not appropriate for partial credit scoring, so will not be discussed further (see Rost, 1988; Andrich, 1978). There are two divide-by-total models appropriate for partial credit scoring: the Generalized Partial Credit Model and the Partial Credit Model, which will be described in detail below.

Generalized Partial Credit Model

When the response categories can be ordered, as in partial credit scoring, Muraki's (1992) Generalized Partial Credit Model (GPCM) can be used. The probability of a response in a particular category (x) for an item (i), where there are $m_i + 1$ categories, conditioned on the examinee's ability level (θ), is defined as:

$$P_{ix}(\theta) = \frac{\exp\left[\sum_{k=0}^{x} a_i(\theta - b_{ik})\right]}{\sum_{h=0}^{m_i} \exp\left[\sum_{k=0}^{h} a_i(\theta - b_{ik})\right]}, \text{ with } \sum_{k=0}^{0} a_i(\theta - b_{ik}) = 0, \quad (6)$$

where a_i is the item discrimination parameter and b_{ik} is the category boundary called the step difficulty parameter for each of the k categories for an item. Figure 6 displays the CRC for an item with 4 categories calibrated using the GPCM.

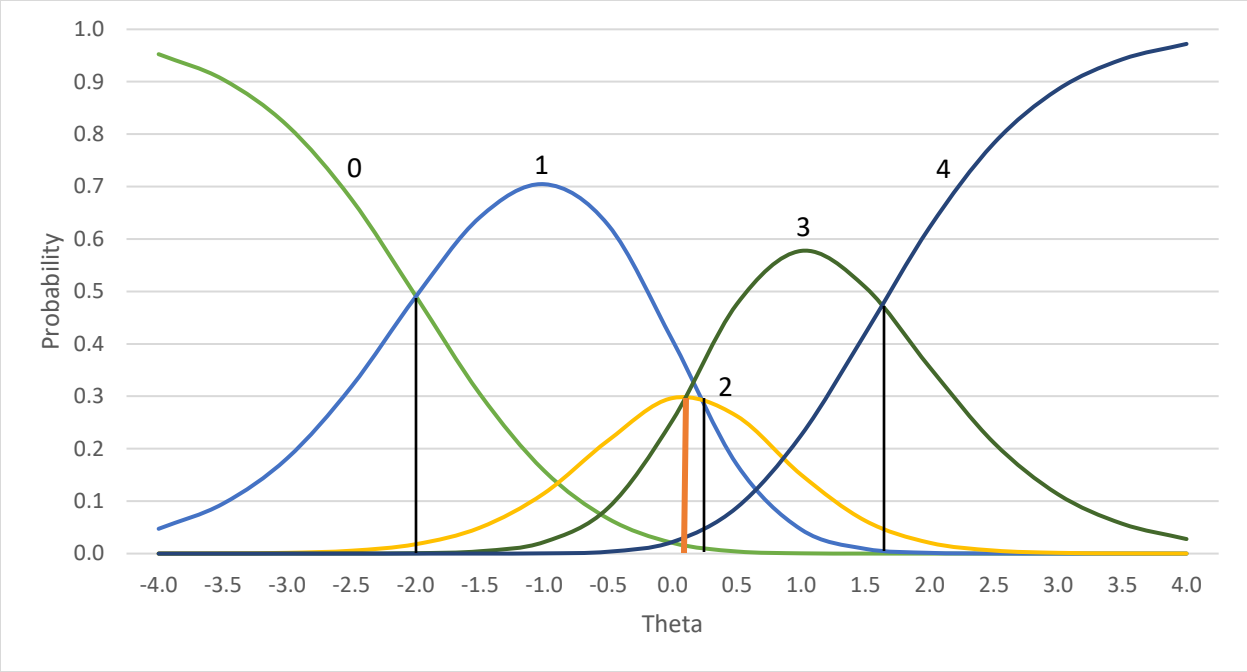

Figure 6. Category Response Curve using the GPCM.

The location of the category boundaries, step difficulties (b_{ik}), indicated by the lines are at the intersection of two adjacent category response curves. Each item is allowed to have differing discrimination parameters and the step difficulties do not have to be ordered, which is seen in Figure 6. The second step difficulty is higher than the third, referred to as a reversal (Dodd & Koch, 1987), with the third step difficulty shown in orange. However, the steps to complete the problem must be completed in order. When items are scored dichotomously, the GPCM simplifies to the 2PL model. Additionally, if all the item discrimination parameters are equal to 1, the GPCM simplifies to the Partial Credit Model (Masters, 1982).

Partial Credit Model

Masters (1982) developed the Partial Credit Model (PCM), which is appropriate for items that have an ordered set of steps to correctly complete the problem which can be scored in a partial credit fashion. Again, like the GPCM, the steps must be completed in order but the step

difficulties need not necessarily be in order of difficulty, unlike the GRM where the threshold difficulties are in order of difficulty. The probability of a particular response (x) in one of $m_i + 1$ categories for an item i, conditioned on the examinee's ability (θ), is defined as:

$$P_{ix}(\theta) = \frac{\exp\left[\sum_{k=0}^{x}(\theta - b_{ik})\right]}{\sum_{h=0}^{m_i}\exp\left[\sum_{k=0}^{h}(\theta - b_{ik})\right]}, \text{ with } \sum_{k=0}^{0}(\theta - b_{ik}) = 0, \tag{7}$$

where, b_{ik} is the point of intersection of probability curves from one category to an adjacent category for m_i categories for an item (i), is defined as the step difficulty parameter. When the discrimination (a) parameters are equal to 1, the GPCM simplifies to the PCM. Additionally, when there are only two categories, the PCM simplifies to the Rasch (1PL) model.

Item and Test Information

Every item, whether dichotomous or polytomous, produces varying amounts of psychometric information, referred to as Fisher's information. Fisher's information indicates the precision of measurement of an item across the range of ability (θ). Fisher's Item Information (Birnbaum, 1968) for a dichotomously scored item i, for a given θ, is defined as:

$$I_i(\theta) = \frac{P'(\theta)^2}{P(\theta)Q(\theta)}, \tag{8}$$

where $P(\theta)$ is the conditional probability of a correct response to item i given theta, $Q(\theta)$ is the conditional probability of an incorrect response to item i or 1- $P(\theta)$, and $P'(\theta)^2$ is the first derivative squared or the slope squared. When items have more than two response categories,

information is calculated with Samejima (1969) general formula for polytomously scored items, with the item information function defined as:

$$I_i(\theta) = \sum_{x=0}^{m_i} \frac{\left[P'_{ix}(\theta)\right]^2}{P_{ix}(\theta)}, \tag{9}$$

where $P_{ix}(\theta)$ is the probability of response in category x for item i conditional on theta, and P'_{ix} is the first derivative. ICCs are the conditional probabilities of a correct response given theta for dichotomous items plotted across the range of theta whereas CRCs are the conditional probabilities of a response in a particular category given theta for polytomous items plotted across the range of theta. Likewise, the information provided by each item, at all points on the ability continuum, can be plotted across the range of theta, producing an item information function as depicted in Figure 7 below.

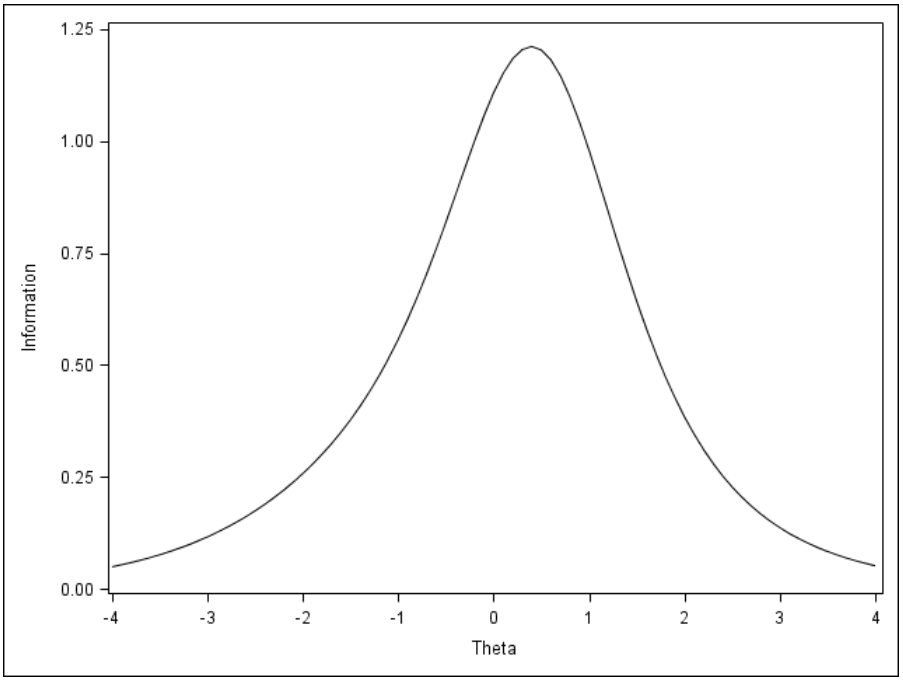

Figure 7. Item Information Function.

An item information function will peak at the ability level (θ) for which the item most precisely measures. The magnitude of the peak for dichotomous items is determined by the discrimination (*a*) parameter. Items with higher discrimination values provide more information for a smaller range of ability levels, producing peaked information functions as seen in Figure 7 (Embretson & Reise, 2000). Lower discrimination values provide less information, or less precise measurement, over a wider range of ability levels, producing flatter item information functions. However, for polytomous items, under the PCM, the magnitude of the peak is related to the order of the step difficulties. Specifically, items that have reversals, as shown in Figure 6, have been shown to produce more peaked item information functions when using the PCM (Dodd & Koch, 1987). The magnitude of the peak of the information function when using the GPCM is driven by the discrimination parameter, producing more peaked functions when this parameter is over 1.0.

An important feature of item information functions is that information functions of items that have been calibrated onto a common scale are additive (Samejima, 1969). The sum of the information functions of the items comprising a test produces test information, which is defined as:

$$TI(\theta) = \sum_{i=1}^{n} I_i(\theta) \ , \tag{10}$$

where $I_i(\theta)$ is the information provided for a given theta (θ) by item *i* for *n* number of items. The information provided by a test can be plotted across the range of ability levels producing a test information function, with the peak occurring at the ability level that is most precisely measured by the test (Embretson & Reise, 2000).

The information provided by a test for a given ability level (θ) is directly related to the precision of measurement at that ability level. Precision of measurement is summarized by the standard error for a given (θ) and is defined as:

$$SE(\theta) = \frac{1}{\sqrt{TI(\theta)}}\ , \tag{11}$$

where the standard error of theta estimate is *SE(θ)*. The relationship between test information for a given theta, *TI(θ)*, and the standard error associated with that ability level is inverse. That is, as information for a given theta value increases, the standard error of the theta estimate will decrease and vice versa.

Computerized Adaptive Testing

Paper & Pencil (P & P) tests are generally designed to measure the average person and, therefore, are composed of many items that are of average difficulty. To achieve accurate measurement, many questions are asked, of which many are too easy or too hard depending on the examinee's ability level. Adaptive tests construct the test so that it is tailored to each examinee's ability level, thereby not wasting the examinee's time with questions that are too easy or too hard. Commonly, CATs start with an item of average difficulty; however, if information about the examinee ability distribution is available, the starting theta (θ) can be based on this prior information. Situations where prior information is not available, starting the test at an average ability level, θ = 0, is a reasonable place. Then, based on the response, correct or incorrect, another item is given that is either harder or easier, respectively. This adaption allows for much shorter tests with much higher precision of measurement, particularly at extreme ability levels (θ). CAT accomplishes this with an algorithm consisting of four basic components:

the item pool, an item selection procedure, an ability estimation method, and a stopping rule (Reckase, 1989).

Item Pool

P & P tests construct different forms of a test based on the premise that they are parallel forms. The items that are selected are from some larger pool of possible items based on a table of specifications, or test "blue print," specifying the number of items, content areas covered, and difficulty of the items. CAT creates parallel forms of a test, however, the forms are not constructed in advance. An item pool is used to select items for examinees specified for their ability. Due to the individualized nature of item selection, the pool needs to be sufficiently large to accommodate the full range of ability (θ) levels. Commonly, large pools are used consisting of hundreds to thousands of items when items are scored dichotomously and exposure control and content balancing procedures are used (Way, 1998).

Each polytomously scored item spans a range of ability levels (θ), therefore, accurate measurement can be achieved with item pools as small as 30 items when exposure control and content balancing are not used (Dodd & de Ayala, 1994). When constraints, such as exposure control and content balancing, are required, as is the case for high stakes testing programs, a much larger pool is necessary (McClarty, Sperling, & Dodd, 2006). It has been suggested that the size of the item pool be based on the length of the test. These items need to be spread across the full range of ability with a similar number of items in the pool at the extreme ability levels (θ) so that precise measurement can be achieved at the extreme theta values. Due to the additive feature of item information, an item pool can be constructed so that the test information function maximizes information at the points on the ability continuum the test is designed to measure

most accurately (Dodd, Koch, & de Ayala, 1993). Although bigger is better, size is not the only consideration for an item pool. The items need to be quality items in order for quality measurement. The parameters specified by the IRT model used to calibrate the items should be within adequate ranges. Specifically, if the 2PL model is used, then the item pool should contain many highly discriminating items in the range of theta the test was designed to measure. Once an item pool of desirable size and characteristics is acquired, the next component of the CAT algorithm is an item selection procedure.

Item Selection Procedures

In general, the item selection procedure for CAT is based on the most current ability estimate obtained using the responses to the previous items. One method is the maximum information method that selects items, conditional on theta, that maximize information or measure the given ability level most precisely (Thissen & Mislevy, 2000). Maximum Fisher item information (MFI; Lord, 1980) is a commonly used item selection method for both dichotomous and polytomous CATs due to the ease of implementation. Fisher's information, as previously defined in the IRT section, selects the item that measures the given ability level most accurately. In other words, it maximizes the information at a given theta. In the unconstrained form, after each item is answered, the interim ability is estimated and the next item selected is the one that provides the most information for that ability level.

Many Bayesian item selection procedures have been developed. The first Bayesian item selection procedure was Owen's Bayesian (Owen, 1975). This procedure selects items that minimize the expected posterior variance of the theta estimate. The item that minimizes the variance of this posterior distribution will be selected for administration. van der Linden (1998)

proposed a variety of Bayesian item selection procedures, including maximum expected information (MEI) and maximum expected posterior weighted information (MEPWI). With these procedures, the expected posterior probability distribution is used to average over the predicted responses for the next item in order to select the item that maximizes the expected information for a given ability level (Choi & Swartz, 2009). Penfield (2006) compared the performance of MEI and maximum posterior weighted information (MPWI; van der Linden, 1998), where the information function is weighted by the posterior distribution, to MFI. The results indicated that the Bayesian procedures produced slightly more efficient estimates compared to MFI. Although many Bayesian procedures exist, they are computationally intensive and produce similar results as the simpler MFI procedure (Choi & Swartz, 2009). Accordingly, the MFI procedure is the most commonly used item selection procedure in CAT and will be used in the proposed study. For more information concerning Bayesian item selection procedures, please see: Owen, R. J. (1975); Pastor, D. A., Dodd, B. G., and Chang. H. –H. (2002); Penfield, R. D. (2006); and van der Linden, W. J. (1998).

There are several considerations before the selection of the most informative item can be made. If item selection is based solely on maximum information, the first few items selected could be the same items for many examinees. Furthermore, if examinees have many items in common those items could quickly become over exposed, or compromised. Constraints on the item selection procedure are needed to limit the exposure of items. Additionally, when multiple content areas are covered in a CAT, content area constraints can be implemented to ensure the specified proportion of items from each content area are administered, which is referred to as content balancing. The items selected should be the most informative given content and exposure constraints. There are several procedures for exposure control, which are classified

into 4 general types: randomization procedures, conditional selection procedures, item stratification procedures, and combinational procedures (Way, 1998).

Randomization Procedures

Polytomously scored items provide more information per item and across a wider range of ability levels compared to dichotomous items (Dodd et al., 1995), which has an impact on exposure control procedures that were originally developed for dichotomous items. Randomization procedures select an item at random from a set of similarly informative items for a particular theta level. Two of the most common randomization procedures used in CAT research with dichotomous IRT models are the Within .10 logits (Lunz & Stahl, 1998) and the Randomesque procedure (Kingsbury & Zara, 1989).

Lunz and Stahl's (1998) within .10 logits, cited in Boyd (2004), selects all items with difficulty parameters (*b*) that are within .10 logits around the current θ estimate and randomly selects one item. This procedure, developed with the Rasch model, selects a set of items based on matching the items' *b*s to the most current estimate of theta, rather than selection based on item information, and continues throughout the length of the test. The Modified Within .10 logits (Davis & Dodd, 2003) selects a set of possible items from which the randomly selected item is chosen. With this procedure, the most informative item that is .10 logits below θ, the most informative item that is .10 logits above θ, and the most informative item at the current θ are selected to comprise the set of items from which one item is randomly selected. This modification was done for the polytomous item extension due to the lack of a single difficulty parameter.

Kingsbury and Zara's (1989) Randomesque procedure selects a set of items based on information. The most informative 5 or 10 items for the current theta estimate are selected, of which one is randomly selected for the dichotomous case. This procedure continues for the entire length of the test in order to decrease test overlap, which is the number of items examinees of similar ability have in common (Kingsbury & Zara, 1989). Davis (2004) modified the Randomesque procedure for the polytomous case to select a set of the most informative 3 or 6 items for the current theta estimate, of which one is randomly selected to be administered.

Conditional Procedures

Conditional item selection is conditioned on a criteria, such as usage, and a parameter is estimated to control the probability of selection. The most commonly used conditional selection procedure is the Sympson-Hetter strategy (Sympson & Hetter, 1985). This is an iterative procedure where the exposure parameter *K* is calculated across a series of simulations, with *K* equaling the probability of the item being administered given that the item was selected. When the value of *K* is high for a particular item, this indicates that this item has not been administered very often and, thus, has a higher probability of being administered if selected. When the value of *K* is low for a particular item, this indicates that the item has been selected and administered often and, thus, would have a much lower probability of administration given that it was selected. This procedure works well at controlling the exposure rate of items; however, it is very labor intensive in that the iterative simulations have to be conducted a priori to estimate the exposure parameter (*K*).

Item Stratification Procedures

Item stratification procedures stratify the item pool according to a statistical property, such as item discrimination (*a*), and then select an item from a particular strata. The first of these stratification procedures to be developed was the *a*-stratified procedure (Chang & Ying, 1999). This procedure was developed to regulate the use of highly discriminating items. When maximum information is used to select items, the highly discriminating items will quickly become over-exposed. Chang and Ying's (1999) procedure regulates the use of these highly informative items by stratifying the pool by the item discrimination parameter (*a*). The items are classified into strata with low *a* values, medium values of *a*, and high *a* values. Additionally, the test is classified into multiple stages: beginning, middle, and end. Chang and Ying (1999) argued that the highly discriminating items are unnecessarily used at the beginning of a test when the interim ability estimate can vary widely. Their solution was to select items from the lower discriminating strata at the beginning of the test and as the test proceeds, items are selected from the more discriminating strata. This leaves the highly informative items (from the high *a* strata) for the end of the test when the ability estimate is not varying widely; therefore, the highly informative items will be used more productively. After the proposition of the *a*-stratified procedure, many variations followed, such as the *a*-stratified with freezing (Parshall, Harmes, & Kromrey, 2000), the *a*-stratified with *b*-blocking (Chang, Qian, & Ying, 2007), multi-dimensional stratification (Lee, Ip, & Fuh, 2002), and the 0-1 stratification strategy (Chang & van der Linden, 2003), just to name a few.

Combination Procedures

The last general type of exposure control procedure is the combination procedure, so named due to the combining of randomization and conditional procedures. The first of these

combinational strategies is the Progressive-Restricted (PR) procedure developed by Revuelta and Ponsoda (1998). This strategy weights the items based on the items' position (*S*) in the test and the information (*I*) provided by that item, calculated with the following equation:

$$W_i = (1 - S)R_i + SI, \tag{12}$$

where W_i is the weight for item *i*, *S* is the serial position which is the number of items administered divided by the total number of items on the test, *I* is the information for item *i,* and R_i is a random number drawn from a uniform distribution. As can be seen, a larger weight is given to the random number at the beginning of the test. As the test continues, the larger weight is given to the item's information, allowing maximum information to have a greater impact toward the end of the test. A major drawback is that this procedure can only be used with fixed length tests or tests with a pre-specified number of items. McClarty, Sperling, and Dodd (2006) developed the Progressive-Restrictive – Standard Error (PR-SE) procedure to extend the application of Revuelta and Ponsoda's (1998) procedure to variable length tests. To accomplish this, the serial position (*S*) is replaced by a ratio of the current SE over the desired SE. This achieves the desired outcome of a larger weight assigned to the random number when the SE is far from the target SE, such as at the beginning of the test. Also, this procedure places more weight on the item information toward the end of the test when the SE is closer to the target.

Research comparing the performance of exposure control procedures assess effectiveness in terms of frequency of use, use of item pool, examinee test overlap, and precision of measurement. Revuelta and Ponsoda's (1998) PR procedure was compared to McClarty et al.'s PR-SE (2006) procedure using Masters (1982) Partial Credit Model. The results indicated that both procedures performed similarly, with the PR-SE increasing item pool utilization (McClarty

et al., 2006). Davis (2004) assessed a variety of methods using the GPCM. This comprehensive study compared the precision of measurement, exposure rate, and difficulty or ease of implementation for the modified within .10 logits, Randomesque, Sypmson-Hetter, conditional Sypmson-Hetter, *a*-Stratified, and enhanced *a*-Stratified procedures. The enhanced *a*-Stratified and the *a*-Stratified performed the worst in terms of measurement precision, exposure rate, and implementation (Davis, 2004). The Sypmson-Hetter and the conditional Sypmson-Hetter procedures achieved the lowest exposure rates; however, this was at the cost of efficiency to implement, item overlap, and pool utilization (Davis, 2004). The Randomesque and modified within .10 logits with 6 item groups were found to be the easiest to implement, while effectively controlling exposure rates (Davis, 2004). Overall, the research indicated that the Rrandomesque or modified within .10 logits with 6 item groups performed the best in terms of pool utilization, exposure rate, test overlap, and ease of implementation.

Content Balancing

Tests that cover multiple content areas require a procedure to ensure that items from each of the content areas are administered according to a pre-specified percentage or test specification (Boyd, Dodd, & Choi, 2010), referred to as content balancing. There are a variety of strategies for content balancing, however, the most commonly used procedure is Kingsbury and Zara's (1989) content constrained CAT (C-CAT). With this procedure, the desired proportions of each content area are first pre-specified. After each item is administered, the proportions of each content area are calculated and compared to the pre-specified proportions. The item with the most information in the content area with the largest discrepancy will be selected to be administered next, given any constraints due to exposure control procedures. Previous research

has shown that this procedure successfully administers specified proportions of items per content area (Boyd, 2004; Davis, 2004; McClarty et al., 2006).

Trait Estimation

Estimation of ability (θ) is calculated using either maximum likelihood or Bayesian methods, with advantages and disadvantages for both. Maximum likelihood estimation (MLE) determines the most likely location of theta by multiplying the probabilities of the individual responses in the response string. A major drawback to this estimation method is that in order to calculate the likelihood distribution of θ, a response in both categories (correct and incorrect) for dichotomous items or a response in two different categories if one response is in either of the extreme categories for polytomous items, is required. After a response in both categories is observed in the response string, the maximum likelihood estimate of θ, $L(\theta)$, is the mode of the distribution.

Until this response string is observed, the decision has to be made as to how the initial ability estimate (θ) should change, which is referred to as step size. Variable step size determines the change in θ based on the range of the items' difficulties within the item pool when content balancing is not used. When content balancing is used, the change in θ is based on the range of the items' difficulties within each content area. This is done to ensure that there is an item within the content area to administer at the θ level. To illustrate in the dichotomous case, if the first response is correct, the next item selected to be administered will be the most informative for an ability (θ) level corresponding to a difficulty (b_i) value that is half the distance to the most extreme difficulty (b_i) value. For example, if the initial ability estimate was θ=0, the first item was answered correctly and the most extreme difficulty is b_i=3.0, the initial ability

(θ=0) estimate would increase to θ=1.5. The next item selected would be the most informative for this ability (θ=1.5) level. If the first question was answered incorrectly and the most extreme difficulty is b_i= -3.0, the initial ability (θ=0) estimate would decrease to θ= -1.5, with the next item selected providing maximum information for this ability level.

In the polytomous case, if the first response was in one of the higher categories, then the initial ability (θ) estimate would be increased to a value that corresponds to half the distance to the highest step difficulty (b_{ik}) value within the content category. Likewise, if the response was in one of the lower categories, then the initial ability (θ) estimate would be decreased to half the distance to the lowest step difficulty (b_{ik}) value within the content category. The next item selected would provide the most information at the new interim ability (θ) estimate. This step size procedure continues until a correct and incorrect response in the dichotomous case, or a response in two different categories if one is in either of the extreme categories in the polytomous case, is observed in the response string.

Previous research with polytomous CATs (Koch & Dodd, 1989; Dodd et al., 1995) demonstrated that the variable step size procedure outperforms the fixed step size procedure, where the change in the initial ability (θ) estimate is a fixed amount. The inability to estimate ability for those examinees that answer all items right or all wrong for dichotomous items, or answer all items in the highest category or all in the lowest category for polytomous items can have a major impact on examinees with abilities in the extreme ranges of theta. For this reason, many use Bayesian methods, which use a prior or known population ability distribution to estimate the probable location of theta.

The Bayes modal estimation procedure uses a prior distribution to determine the most likely location of theta in the posterior distribution. The estimate of theta is the mode of the posterior distribution, as implied by the name. The mean, instead of the mode, of the posterior distribution is used in Expected a Posteriori (EAP) estimation (Bock & Mislevy, 1982), which is the most commonly used in CAT. The prior distribution used, if incorrect, can have an impact on estimation, with a larger impact on shorter tests than longer tests (Mislevy & Stocking, 1989). Previous research with the GPCM has demonstrated that EAP performs similarly to MLE in terms of accuracy of theta estimates when an appropriate prior was used under conditions with similar test length as well as in terms of root mean squared errors under conditions with 20 or more quadrature points used (Chen, Hou, & Dodd, 1998).

Stopping Rule

The stopping rule, also referred to as the termination criteria, is classified into two general types based on the type of test they produce, fixed and variable length (Thissen & Mislevy, 2000). The termination criteria for a fixed-length (FL) test is administration of items until the examinee has been administered a pre-determined number of items. This termination criteria is simple to implement and has the advantage that every examinee completes the same number of items. The drawback is that the precision of measurement will differ across examinees with different ability levels depending heavily on the distribution of items in the pool or the test information function.

Variable-length (VL) tests terminate when a specified precision of measurement (i.e., standard error, SE) is reached. This results in examinees completing different number of items when the SE drops below a specified level of precision, usually SE < 0.3 or 0.2. Examinees are

measured with equal precision, although the length of the test will differ by examinee. A variation of the standard error stopping rule is the minimum information stopping rule, which terminates the test when the items left in the pool to be administered provide such little information (less than a minimum pre-specified amount), that administering more items would be futile. Research using the PCM previously found that the SE termination criteria outperforms the minimum information criteria (Dodd, Koch, & De Ayala, 1993).

Combination procedures merge variable length termination criteria with fixed length criteria. Specifically, the test would terminate when a specified SE is reached or a fixed number of items are administered, whichever occurs first. This combination stopping rule capitalizes on the benefit of equal precision of measurement with the VL criteria, as well as the benefit of efficiency with the FL criteria, stopping the test after a certain number of items so that the test terminates when a precise measurement cannot be achieved.

Adaptive Tests that Allow for Response Review and Revision

All procedures and IRT models discussed thus far adapt at the item level. That is, an item is presented, the examinee responds to that item, an ability (θ) is estimated based on the response, and the next item is selected based on the most current ability estimate. Consequently, the examinee cannot go back and review answers to previous items or change answers to those items. The ability to review previous items and change answers has been shown to decrease anxiety during testing, as well as decrease typographical errors (Lunz, Bergstrom, & Wright, 1992; Stone & Lunz, 1994; Stocking, 1997). High anxiety increases examinee errors during exams, inhibiting an accurate measurement of their ability. The development of adaptive tests has increased test efficiency with shorter tests while increasing measurement precision; however,

this limits examinee flexibility to review and change answers. Wainer (1993) has suggested that allowing examinees to review and change answers will decrease the efficiency of the test, as well as open the door to possible manipulative test taking strategies to bias ability estimates that could be employed, assuming that examinees understand the ability estimation algorithm.

Previous research, which will be described in detail subsequently, has addressed most of these concerns about the impact that response review and revision may have on a CAT that adapts at the item level. Lunz, Bergstrom, and Wright (1992) investigated the impact of response review and revision on test efficiency with licensing and certification exams. Stone and Lunz (1994) expanded this line of research by examining the impact on the ability estimates, test information and precision, as well as decision accuracy with two different examinee populations using two certification tests. Three models were proposed by Stocking (1997) that allow for response review and revision to varying size blocks, or sets, of items. Vispoel, Hendrickson, and Bleiler (2000) assessed the impact of review and revision on the psychometric properties of a vocabulary test in a live testing situation, and assessed examinees' attitudes on review options. Additionally, Vispoel, Clough, Bleiler, Hendrickson, and Ihrig (2002) investigated examinees' ability to distinguish differences in item difficulty in order to bias ability estimates. A new method proposed by Han (2013) addresses the item review and revision issue without restricting revision to after the test is complete. In addition to examining test efficiency and bias, Han (2013) also examined the opportunity to bias ability estimates as suggested by Wainer (1993).

Review and Revision on Licensure and Certification Exams

Early research on the impact of response review and revision utilized licensure and certification exams. Although these types of exams share the same objective as other

educational assessments, to assess the individual's knowledge on a subject and provide an ability estimate, licensure and certification exams are particularly focused on assessing whether or not an individual possesses a minimum competency in the subject area. This minimum competency is commonly assessed by an individual's ability estimate relative to a pass/fail point on the ability continuum.

Lunz, Bergstrom, and Wright (1992) were interested in investigating whether allowing review and revision of responses would substantially decrease the efficiency of the CAT. The efficiency of the CAT was based on the amount of information each item administered provided and the number of items needed to reach a pass/fail decision with a specified level of confidence. The item bank used was constructed from a P & P medical technology certification exam that was field tested on students in medical technology programs across the nation. The items from this P & P test were then calibrated using the 1 PL model. Items that did not fit well were not included in the item bank, creating an item bank of 726 items. The CAT started the test with an item of average difficulty and each subsequent item was randomly selected from the remaining items in the bank that fell within 0.10 logits of the examinee's interim ability level. The stopping rule used was based on a level of confidence, in that the test would stop once the ability estimate was 1.3 times the standard error of measurement above or below the pass/fail cut point. The pass/fail cut point for the exam used was placed at 0.15 logits.

Examinees were randomly assigned to two conditions, one that allowed review and revision of answers (n=220) and a no review condition (n=492). In the review condition, examinees were instructed that after they had completed the exam, they would be allowed to review and revise all of the items, but each item had to be answered when it was first presented. Once the stopping rule was satisfied, the examinees in the review condition were allowed to

review all items, which were presented in the original order with the selected answer highlighted. The examinees in the no review condition were instructed that they had to answer each item as it was presented and that they would have only one opportunity to answer each item. No time constraints were enforced in either condition; therefore, the review condition had unlimited time for review and revision. Additionally, for each examinee in the review condition, two records were maintained, one before review and one after review.

The results indicated that the examinees in the review condition had a slightly higher mean ability (0.24) after review compared to the mean ability of the examinees in the no review condition (0.16). This mean difference in ability between the examinees in the two conditions was statistically significant (t (710) = -2.08, $p < 0.04$). The average number of items administered in the review condition was 96, with an average of 2 items revised. Of the 220 examinees allowed revision, 85 did not revise any items. Of the 135 that revised responses, the maximum number of items revised by one examinee was 16. Among the examinees that revised responses, 30 lowered their ability estimates by revision, 71 improved their estimates, and 34 of the examinees did not change their ability estimates after revision.

Due to review, the efficiency of the test decreased by 1%, on average. The number of additional items needed to recover the information lost during revision depended on the number of items revised. However, for 108 of the 135 examinees that revised at least one item, they would not require administration of any additional items. Of the remaining 27 examinees that revised more items during review, 2-14 additional items would need to be administered to recover the information lost during review. The impact on the pass/fail decision after review was minimal. Only 3 examinees changed the pass/fail decision after review and these examinees' ability estimates before and after review were within one standard error of measurement of the

pass/fail cut point. Examinees whose ability estimates fall very close to the pass/fail cut point have the lowest confidence in the pass/fail decision regardless of whether review is allowed or not. Lunz, Bergstrom, and Wright (1992) concluded that the significant difference in mean ability between the two equivalent groups was due to the review group's ability to correct careless or typographical errors. The majority of the pass/fail decisions did not change and the efficiency that was lost was not substantial enough to support the restriction of review.

Stone and Lunz (1994) extended this line of research by investigating the impact of review and revision by expanding the subjects to two different examinee populations taking two different certification exams. This would allow for any differences in use of review by the different examinee populations and different patterns specific to the test to become apparent. Test precision in terms of information and decision confidence were examined, as well as changes in the pass/fail decision before and after review. Although two different medical technology certification exams were used, review and revision was allowed in both and limited to after the examinee had answered all items.

The study design and methods used were similar to those used by Lunz, Berstrom, and Wright (1992), with the exception of a control group and the addition of another test and examinee population. All examinees took a CAT consisting of a minimum of 50 items and a maximum of 100 items. Two hundred and eight examinees were assigned to take Test 1 and 168 examinees were assigned to take Test 2. Again, all the items for both tests were calibrated using the 1 PL model. Test 1 had an item bank consisting of 664 items and Test 2's item bank was substantially smaller with only 183 items. The stopping rule was increased compared to the previous study, with the test terminating when the examinee's ability estimate was 1.65 times the standard error of measurement, rather than 1.3, above or below the pass/fail cut point. Two

records were maintained for all examinees, one before review and one after review. Based on the records, before review examinees were categorized into low, medium, and high ability levels. The low ability group consisted of examinees with ability estimates more than one SEM below the pass/fail cut point. The medium ability group consisted of those examinees with ability estimates within one SEM above and below the pass/fail cut point. The high ability group consisted of the examinees with ability estimates more than one SEM above the pass/fail cut point. The examinees were also categorized into those that passed and those that failed for both tests before and after review.

The results indicated that the mean ability estimates and standard deviations for both tests increased after review. The average ability estimate for Test 1 increased from .61 to .66 and Test 2's average ability estimate increased from 1.53 to 1.59 after review. The SEM for Test 1 did not change after review, although Test 2's SEM increased slightly from .27 to .28 after review. The information lost due to review, for both tests, could be recovered by the administration of one additional item. Two distinct patterns appeared in pass/fail decisions. Examinees who passed the test before review increased their estimates after review and moved farther above the pass/fail point, thereby increasing the confidence in the pass decision. Examinees who initially failed the test before review increased their estimates after review and moved closer to the pass/fail point, thereby decreasing the confidence in the fail decision. The confidence in the pass/fail decision did not change for those examinees in the high and low ability groups. As expected, it was those examinees close to the pass/fail point where confidence in the pass/fail decision is low. The pattern of revising answers appeared to be random instead of systematic as Wainer (1993) had suggested. Approximately half of the responses were changed from incorrect to correct, which occasionally resulted in a gain in the ability estimate. However, these gains

were sometimes canceled out by changing a correct response to incorrect. Stone and Lunz (1994) concluded, as did Lunz et al. (1992), that the impact of review on the measurement error, ability estimates, and efficiency of the CAT was minimal and did not support the restriction of review.

Stocking Models

Stocking (1997) conducted a simulation study in which three models were proposed and evaluated. The three models that Stocking (1997) proposed provide examinees differing review and revision options, as well as assess the conditional standard error of measurement (CSEM) and the conditional bias across the conditions investigated. All of the CATs simulated used MLE to estimate theta, MFI for item selection, and a fixed length termination criteria. All the conditions simulated examined the worst case scenario of manipulative test taking strategy, or cheating, where it is assumed that any changed answer is changed from incorrect to correct. Additionally, the first few items are answered incorrectly to create the easiest possible test, which likely is not the case in reality. This is referred to as the Wainer Strategy.

The first model (Model 1), simulated examinees were allowed to change answers to a pre-specified number of items after the last item had been answered. Stocking (1997) simulated examinee responses for a 28 item CAT consisting of four conditions with differing number of items allowed for revision: 2 items, 7 items, 14 items, and 28 items. The condition that allowed for two responses to be changed preformed similarly to the conventional 28 item CAT that did not allow for revised responses, with similar CSEM, although the conditional bias for the higher abilities resulted in a gain of about 2 score points. Nonetheless, all other conditions resulted in

large CSEMs and much larger positive conditional bias equivalent to gaining 60 score points for the higher ability levels (Stocking, 1997).

The second model investigated (Model 2) allowed examinees to review and revise any number of items within separately timed sections, to which Stocking (1997) referred to as "Block Review." The content of the items within a block would differ across simulated examinees, but the number of items contained in each block would be constant across simulated examinees. The four conditions for this set of simulations consisted of a CAT: with seven sections of four items, four sections of seven items, two sections of 14 items, or all 28 items in one section. All conditions showed to reduce the CSEM and the positive conditional bias. Specifically, the conditions containing more sections with fewer items per section had CSEM similar to the traditional 28 item CAT with no revisions (Stocking, 1997). Additionally, the positive conditional bias was substantially decreased in the condition with two sections of 14 items each, which resulted in conditional bias equating to an increase of 20 score points. Likewise, the conditions with four sections containing seven items each and seven sections containing four items each both reduced the conditional bias to less than 10 score points (Stocking, 1997).

For the third model (Model 3), simulated examinees could revise answers to items that pertained to a common stimulus. The sets of items, or blocks, were now comprised of items that related to the same stimulus, unlike Model 2 where the content of the blocks of items was heterogeneous. However, items not tied to a stimulus (discrete items) could not be revised. Again, four conditions were investigated each using a different item pool: a CAT consisting of 28 items selected from an item pool containing two blocks with four items each; a CAT with 30 items selected from an item pool containing three blocks with eight items each; a CAT with 35 items selected from an item pool containing six blocks with 26 items each; and a CAT with 31

items selected from an item pool with seven blocks containing 31 items (Stocking, 1997). All conditions had similar CSEM and conditional bias compared to the traditional, no revision CAT. Still, this model was limiting in the sense that discrete items could not be revised and items within a set could not be skipped. All of the models Stocking (1997) investigated had some form of restricted review options; however, skipping items to return to later was strictly not allowed. Results indicated that only the most restrictive conditions achieved CSEMs and conditional biases within acceptable ranges, as well as robustness to the Wainer Strategy.

Review and Revision on CAT Vocabulary Tests

Results from previous studies that suggested a minimal loss in efficiency and measurement accuracy with review and revision lead Vispoel, Hendrickson, and Bleiler (2000) to examine response review and revision on vocabulary CAT tests, as well as examinees' desire for review options. Although a majority of Stockings' (1997) models resulted in biased ability estimates, the study was a simulation in which the human element is eliminated and the worst case scenario of cheating was simulated. Vispoel et al. (2000) designed a live testing study to examine some of the restricted review options Stocking (1997) used in order to gain a better understanding of review behavior among real examinees.

A convenience sample of 242 participants from the University of Iowa Introductory Educational Psychology and statistics courses volunteered for the study. Each student completed a test anxiety inventory, a fixed length 40-item vocabulary skills CAT, and a questionnaire including demographic information and attitudinal questions about tests. The vocabulary test was constructed from an item pool of 609 items. The participants were randomly assigned to four conditions: full review of all items at the end of the test, no review, and two forms of block

review. The block review conditions allowed for review of items within a block or set after completion of the block of items and the CAT was designed to adapt both within and between blocks. The block review conditions were eight blocks of five items or four blocks with ten items. Skipping items was not allowed, but items could be marked for later review. Still, once an examinee moved to a new block, the previous blocks could not be reviewed or revised. The items were calibrated using a modified 3 PL model that freely estimated the difficulty and discrimination parameters but fixed the pseudo-guessing parameter at 0.15. Item selection was based on maximum information, with no exposure or content constraints. Bayesian EAP estimation was used to estimate examinee ability based on previous research (Vispoel et al., 1999), suggesting that EAP is less susceptible to score distortion due to the Wainer strategy seen in the worst case scenario of Stocking's (1997) study in which ML estimation was used.

The results indicated, as did previous research, that 47.5% of examinees in the review conditions changed answers to at least one item, with more answers changed from wrong to right. However, the percentage of items revised only comprised 2.31% of the overall items administered. The majority of examinees in the review conditions that revised answers improved their ability estimates after review and revision, although measurement precision changed very little after review with a precision ratio of .991. A positive relationship was found between block size, number of items marked for review, number of answers revised, and time spent on reviewing answers, with the latter three increasing when block size increased. Approximately 96% and 95% of the examinees indicated that answer review and question marking options are desirable in CATs, respectively. Additionally, examinees reported that marking answers for later review as their most commonly used test taking strategy. Again, no evidence of the Wainer strategy was supported by the results and Vispoel et al. (2000) concluded that allowing limited

review options would increase ability estimate validity with minimal impact on measurement precision and test efficiency.

Since Wainer (1993) suggested that examinees could devise a strategy to bias ability estimates if response review and revision was allowed, most studies on review options have examined the plausibility of this strategy and found it not plausible. Vispoel, Clough, Bleiler, Hendrickson, and Ihrig (2002) took a closer look at examinees' ability to distinguish differences in item difficulties. Vispoel et al. (2002) taught the participants two different strategies, the Kingsbury and the Generalized Kingsbury strategy, to explore the possible bias in ability estimates due to these manipulative strategies. Kingsbury (1996) described the Kingsbury strategy in a paper presented at the National Council on Measurement in Education annual meeting. This strategy is based on knowledge of the item selection algorithm in which an item answered correctly will result in a harder item subsequently administered and, likewise, an item answered incorrectly will result in an easier item subsequently administered. The examinee would mark an item for review if they were unsure of their answer if the next item presented was easier, indicating that the response to the previous question was incorrect. However, this strategy assumes that examinees can distinguish item difficulties in pairs of items and that examinees only utilize this strategy when they are unsure of their answer. The Generalized Kingsbury strategy, discussed in Wise, Finney, Enders, Freeman, and Severance (1999), eliminates the second assumption that examinees only use the strategy when they are unsure of their answer, but rather use it for every item.

Vispoel et al. (2002) expanded upon the design of Vispoel et al. (2000) with the addition of the two testing strategies using the same vocabulary test and two review conditions. The vocabulary tests utilized the same item pool, item selection, and ability estimation procedures as

Vispoel et al. (2000). The participants were randomly assigned to one of the seven conditions: no strategy/no review (NR), no strategy and review of all 40 items after completion of all items (R40) , no strategy and review of eight blocks of five items after completion of the block (R5), Kingsbury strategy and review of all 40 item after completion of all items (K40), Kingsbury strategy and review of eight blocks of five items (K5), Generalized Kingsbury strategy and review of all 40 items (GK40), and Generalized Kingsbury strategy with review of eight blocks of five items (GK5). The no review and the two no strategy conditions served as baseline and replication conditions to compare to Vispoel et al. (2000). The participants assigned to the strategy conditions were taught the two testing strategies and given an opportunity to practice applying them before starting the vocabulary tests.

Vispoel et al. (2002) examined the consistency of the item selection procedure in adhering to correct answers leading to a harder item and incorrect answers leading to an easier item algorithm, due to the testing strategies basis in this algorithm. Because maximum information was used as the item selection procedure, it was expected that the second half of the CATs would depart more from the algorithm due to less discriminating items remaining in the item pool for selection. The results, measured by the proportion of items following the algorithm, supported this expectation with consistency to the strict correct-harder item, incorrect-easier item algorithm found in 88% of the first 20 items across the whole sample and a drop in consistency to 73% in the last 20 items. Examinees' ability to distinguish differences in item difficulty was slightly greater than chance at an average of 0.61, consistent with results from Wise et al. (1999). The item pool used in Vispoel et al. (2002) contained fewer items with absolute *b*-values greater than 0.5 logits apart, only 13% of the items in the pool had absolute *b*-values greater than 0.5 logits apart. For those items with absolute *b*-values greater than 0.5 logits

apart, the success of examinees to distinguish the differences in difficulty only increased to 0.67 from 0.61, on average, which was lower than 0.73 as found in Wise et al. (1999). Vispoel et al. (2002) proposed that this decrease in success for distinguishing differences in pairs of items is due to differences in the distribution of item pools used in the two studies.

As was expected, testing time increased for the review conditions compared to the no review conditions, with an 11% increase for the R5 condition and 20% increase for the R40 condition. Additionally, the testing strategy conditions saw a bigger increase in testing time with the GK40 resulting in the largest increase in testing time (a 52% increase) to complete and review all items. The results concerning the ability estimates support previous findings in that item review slightly improves ability estimates with a mean increase in the review conditions of 0.03. Interestingly, the two testing strategies, Kingsbury and Generalized Kingsbury, both decreased mean ability estimates after review with a mean of -0.04 and -0.07, respectively. This result was not particularly expected and provides evidence that the use of these strategies will hurt examinees estimates rather than inflate the estimates as Kingsbury (1996) and Wise et al. (1999) originally proposed. In the review conditions, the answer changing behavior followed previous patterns, with more answers changed from wrong to right. However, this pattern reversed with the strategy conditions, with more answers changed from right to wrong.

These results provide evidence that the two test strategies studied would likely not provide examinees any advantage. Rather, test taking strategies would hurt examinees' ability estimates. The lack of examinees' ability to distinguish differences in difficulties between pairs of items in combination with the item selection algorithm's inconsistency, especially in the second half of the test, produces a potentially detrimental strategy. It should also be noted that this study did not use any exposure control or content balancing procedures, which would likely

increase the item selection algorithm's inconsistency, leading to a greater detriment to examinees' ability estimates. Further, the increased testing time required to implement these strategies would be prohibitive under high stakes testing situations.

Item Pocket Method

Han (2013) developed a new method, called the item pocket (IP) method, to allow for greater flexibility on the examinees part to review, revise, and skip items in a CAT that adapts at the item level. Han (2013) argues that although Stocking's Model 2 performed well in terms of conditional bias and the CSEMs when there were more separately timed sections containing fewer items, it did not allow examinees to skip items, which could have an impact on test efficiency. Although examinees can revise answers within a section, they must answer each question first, and could resort to randomly selecting an answer in order to move forward will inevitably decrease the efficiency of the CAT because examinees' random responses are used to select subsequent items, which may not reflect their true ability. Han (2013) also argues that small separately timed sections may not be realistic for an operational testing program, which the IP method addresses.

The IP method, proposed by Han (2013), creates a "pocket" in which examinees can place items at any time during the test. The items placed in the pocket are not used in the item selection algorithm, so restrictions implemented in Stocking's (1997) Model 1 and 2 and in the work done by Vispoel et al. (2000 & 2002) are no longer necessary. The items in the pocket can be reviewed at any time during the test and once a final answer is confirmed, the item is removed from the pocket. This method allows examinees to revise all items, if time limits allow, which

provides greater flexibility to move through the test without jeopardizing the efficiency of the CAT algorithms.

Han (2013) assessed this new method in terms of CSEM and conditional bias, with simulated examinee responses to a fixed length, 40 item CAT. The simulation study design included items from an operational CAT item pool calibrated with the 3PL model, using MFI as the item selection criteria with the Sympson-Hetter (1985) exposure control procedure and MLE to estimate ability. Four IP size conditions were investigated, including zero items (a baseline, conventional CAT), two items, four items, and six items. Additionally, any items left in the pocket would be counted as incorrect under the IP design. Nonetheless, the simulation study had no time restrictions, meaning that no items were left in the pocket, which Han (2013) admits does not reflect realistic testing conditions.

Due to the simulated nature of the study, items were selected to be placed in the pocket based on the discrepancy between the examinee's known ability (θ) level and the item's difficulty. This discrepancy between the examinee's known ability and the item difficulty was used to simulate which items the examinees would find difficult. Only items that had a difficulty higher than the examinee's known ability would be selected for placement in the item pocket. If the item's difficulty (*b*) was half a theta unit higher than the examinee's known ability, then the item was deemed challenging and placed in the pocket 70% of the time. If this discrepancy was less than 0.5, then the item was deemed challenging 50% of the time and placed in the pocket.

In situations where the pocket was full and the current item is to be placed in the pocket, the simulated examinee would compare the items in the pocket to determine the easiest item to remove so that the current item can be placed in the pocket. To determine the easiest item, all

pairs of items in the pocket are compared. Those items with discrepancies greater than 0.5 logits higher than the examinee's known ability, the easiest item was simulated as being selected 70% of the time. If the discrepancy was between the examinee's known ability and 0.49 logits higher than the ability, the easiest item was simulated to be selected 50% of the time. Once the easiest item in the pocket was identified, that item was compared to the current item, with discrepancies greater than 0.5 logits higher than the known ability simulated as being selected 70% of the time. Again, when the discrepancy is between the known ability and 0.49 logits higher, the easiest item would be simulated being selected 50% of the time. Once the item that is to be answered is selected, either the item removed from the pocket or the current item based on the above described comparisons, the item is administered.

Han (2013) selected these percentages of determining placement in the pocket and selection of the easiest item within the pocket based on research by Wise et al. (1999) and Vispoel et al. (2002) on examinee test taking strategies when the examinees have an opportunity to review and change answers. Vispoel et. al. (2002) found that examinees are not accurate in determining the most difficult item when comparing pairs of items, with accuracy increasing as the difference in difficulty between the two items increases. In an attempt to more closely simulate examinee testing behavior, these percentages introduce error in determining the items that are placed in the pocket and selecting the easiest item in pairs of items. The computer can select the easiest item every time, however, Vispoel et al. (2002) demonstrated that examinees are not very successful in determining item difficulty.

Results indicated improved robustness to positive conditional bias, as seen with Stocking's (1997) Model 1 (with only two revised items) and Model 2 with four or more separately timed sections. Specifically, comparing the conditional bias of the θ estimates within

the range of -2 to +2 for IP size 2, 4, and 6 to the baseline condition showed either no change for the two and four IP size conditions, and slight positive bias for the lower ability estimates and slight negative bias for the higher ability estimates with an IP size of six (Han, 2013). The CSEM showed a slight increase in the θ estimates of less than 0.10 across the theta range of -2.5 to +2.5 for all of the item pocket sizes.

The possible use of manipulative test taking strategies to improve scores was also assessed, due to the concern in previous research in examinee cheating. The design of the IP method excludes items placed in the pocket from use in the item selection algorithm, meaning the items placed in the pocket are not used to select the subsequent items. The strategy suggested by Wainer (1993), where examinees purposely answer initial items incorrectly in order to get subsequently easier items, thereby artificially increasing scores, is not possible. The exclusion of items placed in the pocket from the item selection algorithms eliminates the use of both the Kingsbury and Generalized Kingsbury strategies, as well. The impact of using the pocket on test completion was not directly assessed since there was no time limit and all simulated examinees responded to all 40 items. Han (2013) suggests that in operational testing programs, where some time limit is placed on examinees, IP usage for the lower ability examinees could inhibit the completion of the test. The size of the pocket, suggested by Han (2013), should be based on the length of the test, limiting it to 20% of the test. The IP method displayed robustness to manipulative testing strategies while maintaining efficiency and precision, as well as providing more flexibility to examinees.

Statement of Problem

The development of CAT has increased assessment efficiency, while also increasing measurement precision. The increased efficiency has decreased the demand on examinees; however, the adaptive nature of the tests has restricted examinees' control in moving through a test as they would with a P & P tests with the opportunity to skip questions, as well as review and change answers. This restriction was necessary due to the ability estimation algorithm, which is estimated after each item based on the response to that item. Allowing examinees to change answers could open the door for cheating.

For instance, the Wainer strategy is the purposeful answering of items incorrectly, thereby creating an easy test. When the examinee is allowed review and revision, the examinee goes back through this artificially easy test and answers all the items correctly, resulting in an inflated score. The Kingsbury and Generalized Kingsbury strategies use the information from the subsequent item, or examinees' perception of the item's difficulty, to gage whether the previous item was answered correctly. If the subsequent item is more difficult, then the previous item was answered correctly. If the subsequent item is easier than the previous item, the examinee can assume the response to the previous item was incorrect. The examinee uses this information to correct the items that were answered incorrectly. Previous research by Lunz et al. (1992), Stone and Lunz (1994), and Vispoel et al. (2000 & 2002) provides evidence that the Wainer strategy and both the Kingsbury and Generalized Kingsbury strategies are not plausible and are not likely to be used by examinees in either low or high stakes testing.

Although relatively little research has been conducted on anxiety from CAT restrictions, the restrictive testing procedures could have the effect of increasing examinee anxiety, which

could result in poor measurement of those examinees. It has also been suggested that disallowing review could increase measurement error due to typographical errors that examinees would have caught had they had the chance to review their item responses (Lunz et al., 1992; Stone & Lunz, 1994; Stocking, 1997).

Stocking (1997) extended previous research by Stone and Lunz (1994) with the development of three models that allowed restricted revision options. However, these restrictions were limiting and conditional bias at extreme theta levels was found mostly out of the acceptable range for operational testing programs. Han's (2013) IP method provides a viable option to address examinees' control in moving through the test, while demonstrating robustness to cheating and conditional bias within acceptable theta ranges. Han's (2013) research was conducted using a dichotomous 3PL model; however, most operational testing programs, such as the Scholastic Achievement Test (SAT), contain both dichotomous and polytomously scored items. Before an extension to mixed format tests is examined, extension of Han's (2013) IP method to the polytomous case is needed. Currently, no research has been conducted to date that allows for review, revision, or skipping questions using a polytomous IRT model.

A limitation of Han's (2013) study was the lack of content balancing in the simulation design, which does not reflect existing operational testing programs. Han's (2013) study employed one termination criteria, which was a fixed number of items. Therefore, the investigation of a variable length termination criteria will contribute to the applicability of this new method to a wider variety of adaptive assessments. In addition, test length was not varied in Han's (2013) simulation study, thus, varying test length should be explored to determine the impact test length has on measurement precision in conjunction to the IP method. Han's (2013) simulation study employed no time limit and therefore no items were remaining in the IP. As

such, two potential outcomes for items remaining in the IP at the conclusion of the test, forced to answer or ignoring them, should be examined to determine the impact on measurement precision. The forced answer (FA) condition will replicate the condition in Han's (2013) study, whereas the ignore (Ign) condition will simulate the impact on precision of measurement when the items are disregarded as if the examinee never saw them.

Research Questions

The purpose of this dissertation research is to investigate the performance of a CAT using the IP method with polytomously-scored items that are calibrated using the GPCM. The impact of different item pocket sizes and termination criteria will also be evaluated. In addition, the performance of a CAT using the IP method will be compared to a baseline CAT without implementing the IP method. Four main research questions will be addressed in this study:

1) What is the impact of the IP method on precision of measurement across the range of ability levels when applied to a CAT using the GPCM with content balancing and exposure control procedures?

2) What is the impact on precision of measurement under the two termination criteria (i.e., fixed and variable length)?

3) What is the impact of the two item completion conditions (forced answer or ignored) on precision of measurement?

4) What impact does implementing the IP method have on test efficiency in the variable length conditions?

Chapter 3: Methodology

Design Overview

The CAT simulation study extended the application of the IP method to a polytomous CAT using the Generalized Partial Credit Model (GPCM). The application of this method to a polytomous model was evaluated in terms of IP usage, conditional bias, precision of measurement across the range of ability (θ), and administration efficiency.

The application of the IP method to the GPCM was investigated under three pocket size conditions, including two, three, and four items, compared to a baseline condition without the application of an IP. Han (2013) investigated three pocket size conditions of two, four, and six items; however, his study employed a dichotomous IRT model which generally requires longer tests than tests using polytomous items for more accurate person and item measurement. It has been shown that polytomous items provide more information per item (Koch & Dodd, 1989) and, thus, shorter tests can achieve accurate person and item measurement. Han (2013) suggested that the IP size be based heavily on the length of the test, with the pocket containing no more than 20% of the items on the test. As such, smaller IP sizes were chosen for this study because polytomously-scored items will be used with 15- and 20-item tests.

Han's (2013) study investigated only a fixed length stopping rule. The termination criterion in CATs may impact the precision of measurement. Accordingly, the current study used two stopping rule conditions, two fixed length condition and variable length conditions , in which administration will stop when a specified precision of measurement ($SE \leq 0.3$) has been achieved or the maximum number of items has been administered. Han's (2013) study employed only one test length, 40 items; thus, the current study used two test length conditions, 15- and 20-items tests.

Under Han's (2013) design, no items were left in the item pocket at the completion of the test, meaning that before the test would conclude, the items in the pocket had to be answered. The current study investigated the impact on measurement precision under two item completion conditions: the items in the pocket are ignored (Ign) or examinees are forced to answer (FA), matching Han's (2013) design.

In sum, four independent variables were manipulated, including IP size (2, 3, 4), item completion design (ignore items in the pocket and forced completion of items), test length (15 and 20 items), and CAT stopping rule (fixed-length and variable-length), resulting in a completely crossed 3 x 2 x 2 x 2 factorial design with 24 conditions. In addition, four baseline traditional CAT conditions in which the IP method is not implemented were included, resulting in 28 total conditions. All conditions implemented content balancing using a content constrained CAT (C-CAT; Kingsbury & Zara, 1989) and exposure control using Kingsbury and Zara's (1989) Randomesque procedure with a six item group size. Ability estimation utilized Maximum Likelihood Estimation (MLE) with a variable step size adjustment implemented until an ability estimate could be obtained. Each condition had 1,000 simulated examinees sampled from a normal distribution with 500 replications.

Item Pool and Test Characteristics

The item pool that was used in this study is based on a national testing program, consisting of 157 constructed response items. This pool includes items in three content areas, with the first content area (I) containing 61 items, content area II containing 59 items, and content area III having the fewest with 37 items. Within each content area, items differ in the number of steps to a solution or number of response categories. Each content area includes items

with three, four, or five response categories which corresponds to the possible number of score points of two, three, and four, respectively, for each item. The item pool contains 99 three-category items, 29 four-category items, and 29 five-category items. The item parameters from this national testing program item pool are the same as was used in Davis' (2004) study. Table 1 displays the percentage of items by content area and number of response categories. According to Davis (2004), this item pool's test information peaks at $\theta = -0.6$. Descriptive statistics for this item pool are shown in Table 2 and provide the mean item discrimination and step values across the three content areas (Davis, 2004).

	Content Areas		
No. of Categories	**Area I**	**Area II**	**Area III**
3	24.57%	23.63%	14.81%
4	7.22%	6.94%	4.35%
5	7.23%	6.94%	4.35%

Table 1. Percentage of Items by Content Area and Number of Response Categories

	Discrimination	**Step Difficulty 1**	**Step Difficulty 2**	**Step Difficulty 3**	**Step Difficulty 4**
Mean	0.92	-0.99	0.18	-0.19	-0.12
SD	0.19	0.90	0.99	0.76	0.90
Minimum	0.54	-3.13	-1.81	-1.48	-2.36
Maximum	1.52	1.50	3.57	1.51	2.34
n	157	157	157	58	29

Table 2. Descriptive Statistics for Item Parameters for Item Pool

Data Generation

To simulate a population of examinees whose ability distribution matches the item pool, 1,000 examinee ability (θ) levels were randomly selected from a normal distribution with a mean

of 0 and a standard deviation of one, with 500 replications. Examinee item responses were generated based on the GPCM using the IRTGEN SAS program developed by Whittaker, Fitzpatrick, Williams, and Dodd (2003). To generate item responses for each examinee, the program first calculates the probability of responding in each category using Equation 6. Then, the probabilities for each category for an item are summed to provide a cumulative subtotal. This cumulative subtotal is compared to a random number drawn from a uniform distribution. If the random number is less than or equal to the cumulative subtotal for a particular category, then the examinee is assigned that category score for that item. This process continues for all simulated examinees for all items in the pool.

CAT Simulation

The CAT simulation used Davis' (2004) constrained CAT program to estimate examinees ability (θ) level using the GPCM with modifications to include an IP. For each condition, examinee responses and item pool characteristics were input into the program for simulation with item selection utilizing the program's default algorithm, Maximum Fisher item information (MFI).

All simulated examinees began the test with an initial ability (θ) estimate slightly above the population distribution mean, $\theta = 0$. Before the interim ability can be estimated with MLE, a modified variable step size, within a particular content area, was used to adjust the initial ability (θ) estimate. This continued until a response in two different categories, if one response was in either of the extreme categories, was achieved, after which MLE provides interim ability estimates for the remainder of the CAT.

Item selection based on MFI was constrained with the Randomesque exposure control procedure in all conditions. Content balancing was done using content constrained CAT (C-CAT) based on the joint proportions of content area and number of categories (see Table 1) in all conditions. Based on Davis' (2004) research, the Randomesque procedure performs optimally when the six item group size is used with polytomous items. Therefore, for all conditions, the six item group size was used in the present study. The C-CAT content balancing procedure began the test by randomly selecting a content area from which the first item was selected. After the initial item was selected, the procedure iteratively compared the joint proportions in Table 1 to determine the content area with the largest discrepancy between the target proportions and administered item proportions. The content area with the largest discrepancy had an item selected for administration.

Simulation of examinee usage of the IP followed the procedure used by Han (2013). Han assumed that examinees will place items that they find challenging in the pocket to come back to later if time allows. An item placed in the pocket will only be used for the C-CAT balancing procedure and excluded from the ability estimation until the answer is finalized, at which point that item is removed from the pocket. Similar to Han's (2013) simulation study, a procedure was used to determine which items are selected for placement in the pocket in the current simulation study. Han (2013) simulated which items the simulees would find difficult and therefore placed in the pocket by comparing the simulated examinee's known ability level (θ) to the item's difficulty (b) parameter. When the known ability level was more than 0.5 logits below the item's difficulty, the examinee would be designated as finding the item challenging. Han's (2013) study used the 3PL model, in which each item has one difficulty (b) parameter. Polytomous items cover a range of ability levels, meaning each item has multiple b parameters or step

difficulties. However, the information functions for polytomous items peak at the point on the ability scale (θ) where the item provides the most information about an examinee at that ability level. Polytomous items can be selected based on information functions (Kamakura & Srivastava, 1982). For this study, an item was considered challenging for a simulated examinee if the examinee's known θ was 0.5 logits below the ability level that corresponds to the peak of the item's information function. Only items in which the peak of their item information function was found to be above the simulated examinee's ability level were selected for placement in the item pocket.

When the examinee's true ability was located below the ability level indicated by the peak of the information function by 0.49 logits or less, the item was selected for placement in the pocket 50% of the time. When the examinee's known ability level was below the ability level indicated by the information function by 0.5 logits or more, the item was selected for placement in the pocket 70% of the time. When an item met these two conditions and the IP was full, the current item and the item(s) in the pocket were compared to determine if an item would be removed and administered or the current item would be administered. The easiest item in the paired comparison would be selected based on the discrepancy between the two items' ability levels corresponding to the peaks of the items' information functions. When this discrepancy was 0.49 logits or less, the easiest item was selected 50% of the time. When the discrepancy between two items was 0.5 logits or greater, then the easiest item was selected 70% of the time. Based on these comparisons, the current item would be answered if it was deemed easiest or the easiest item in the pocket would be answered to make room in the pocket for the current item. Once the item is selected for administration it was administered and the response for that item based on the simulee's known ability level was recorded.

The percentage in determining the easiest item to answer was based on research by Vispoel et al. (2002) in a live testing situation on test taking strategies and examinee's accuracy in determining the difficulty of items. Vispoel et al. (2002) found that examinees are not accurate in determining the most difficult item when comparing pairs of items, with accuracy increasing as the difference in difficulty between the two items increases. Although examinees are not very accurate in selecting the easiest item in pairs of items, the computer can always select the easiest item. Therefore, the percentages were used to introduce error in determining the items placed in the pocket and in determining the easiest item in the pocket to answer.

Han's (2013) study required all items in the pocket to be answered before concluding the test. The forced answer (FA) item completion condition in the current study simulates the same requirement used in Han's study. An additional item completion condition was added to the current study in order to examine the impact on measurement precision when items in the pocket are ignored at the conclusion of the test. In this situation, those items are simply disregarded, not administered and the test concludes. Additionally, those items placed in the pocket are above the examinee's known ability level and could have been placed in the pocket toward the beginning of the test, therefore, would likely not provide very much information. The forced answer (FA) condition does confound the stopping rule, however, which will be discussed further below.

All simulated CATs terminated under two termination criteria: fixed or variable length stopping rules. Additionally, all simulated CATs with an IP applied two item completion conditions: (1) forced answer (FA) and (2) ignored (Ign). Further, the fixed length tests terminated once the examinee had completed either 15 or 20 items. Test length has a direct impact on measurement precision, in that as test length increases measurement precision increases, although, examinees are not measured with the same precision across the ability

continuum. Specifically, the examinees' with extreme abilities, either low or high, will be measured with less precision. The fixed length criteria of 20 items was based on the comprehensive study by Davis (2004) in which the same item pool and polytomous IRT model was utilized. The 15 item criteria was included to investigate the impact of precision of measurement with a shorter test. The forced answer (FA) condition is meant to replicate Han's (2013) study by extending findings to the polytomous case. However, because the fixed length will be compared to the variable length conditions, some adjustments were made. Specifically, in the forced answer (FA) item completion condition, the simulated examinees are required to answer item(s) in the pocket, as was the case in Han's (2013) simulation. Therefore, depending on the number of items in the pocket, the examinee is forced to answer the item(s) once the termination criterion has been met. For instance, if the number of items in the IP is three, the examinee will have to answer those three items after the 15 or 20 items have been administered, with 18 or 23 total items administered, respectively. In the ignored (Ign) item completion condition, the simulees ignore the item(s) left in the IP and are administered 15 or 20 items total, depending on the test length condition.

The variable length stopping rule terminated the test once a pre-specified precision of measurement was achieved (i.e., $SE \leq 0.3$) or the maximum number of items was completed (15 or 20 items), whichever came first. The variable length termination criteria has the advantage of shorter tests and equal measurement precision for those examinees' whose abilities are matched to the item pool distribution. Meaning the test was designed to measure those abilities with many informative items for those abilities included in the item pool. Again, under the forced answer (FA) item completion condition, the examinee was forced to answer the item(s) in the IP after the SE dropped below the termination criteria (0.3), or the maximum number of items had

been administered. For instance, for an examinee with three items in the IP and if the SE dropped below the 0.30 criteria at item 12, the examinee would be forced to answer the three items in the pocket, resulting in a total of 15 items administered. Conversely, for an examinee with three items in the IP and if the SE never drops below the 0.30 criteria, this examinee would be forced to answer the three items in the pocket, resulting in a total of 18 items administered. This modification allows for the variable length conditions to be comparable to the fixed length conditions. In situations where the variable length SE criteria is met, the items in the pocket are answered, resulting in two, three, or four more items administered after the termination criteria is met in the two, three, and four item IP size, respectively. Therefore, the fixed length conditions must follow this same criteria, administering the items in the pocket after the termination criteria is met, resulting in two, three, or four more items administered depending on the IP size condition. Under the ignore (Ign) item(s) completion condition, once the termination criteria was satisfied, fixed or variable length, the test terminated, ignoring the item(s) remaining in the pocket.

Data Analysis

The results of the simulated CAT using the item pocket method was analyzed in terms of (1) item pocket usage, (2) the overall precision of measurement of the final theta estimates (including mean conditional standard error of the ability estimates, mean bias, and root mean square error), and (3) test efficiency. The use of MLE in ability estimation may lead to nonconvergent cases. In these cases, MLE is not implemented or the final θ estimate is below -4 or above +4 (Gorin, Dodd, Fitzpatrick, & Shieh, 2005). These nonconvergent cases were listwise deleted in all conditions before the outcome measures were calculated. However, for each condition, the mean number of nonconvergent cases across the 500 replications in a

condition as well as the minimum and maximum number of nonconvergent cases within a condition are reported.

The overall precision of measurement was assessed in a variety of ways. Descriptive statistics of the final θ estimates and their standard errors for each condition are reported, as well as the grand mean, mean minimum, and mean maximum theta value per condition. The impact of the two item completion conditions on the final theta estimates is of interest. The impact was evaluated in terms of the overall precision of measurement. Descriptive statistics for the two item completion conditions of the simulees' final θ estimates and their standard errors are reported as well as the grand mean, mean minimum, and mean maximum across the 500 replications within these conditions. Recovery of the known thetas was evaluated using the mean Pearson product moment correlation across the 500 replications per condition, as well as the minimum and maximum correlation between the known and estimated θ values in a condition. Theta recovery was also evaluated using bias and root mean square error (RMSE). Bias assesses the systematic error of measurement in the final theta estimates and is defined as:

$$Bias = \frac{\sum_{k}^{n}\left(\hat{\theta}_k - \theta_k\right)}{n}, \tag{13}$$

where $\hat{\theta}_k$ is the final theta estimate for simulee k, θ_k is simulee k's known theta, and n is the number of simulees. Bias was averaged over the 500 replications and plotted across the range of theta with 0.5 increments for the three IP size conditions. RMSE assesses the total error of measurement and composed of bias and standard error in the final theta estimates. RMSE is defined as:

$$RMSE = \sqrt{\frac{\sum_{k}^{n}\left(\hat{\theta}_k - \theta_k\right)^2}{n}} \quad , \tag{14}$$

where $\hat{\theta}_k$ is the final theta estimate for simulee *k*, θ_k is simulee *k*'s known theta, and *n* is the number of simulees, which was averaged over the 500 replications per condition. The CSEM assesses the precision of measurement at different ability levels, θ. For all conditions, the standard error of measurement of final theta estimates with 0.5 increments across the range of θ was averaged over the 500 replications, producing the grand mean CSEM. These were plotted across the ability scale for the IP size conditions, producing conditional plots to assess the precision of measurement in the final theta estimates.

The overall IP use was assessed with descriptive statistics, including the mean, minimum, and maximum IP use across the 500 replications in each IP size condition. The IP usage was also assessed conditionally on θ with 0.5 increments across the range of θ because it is accepted that the use of the IP will vary by examinee ability level. Therefore, for each condition, the grand mean of IP usage was calculated, conditional on known theta, by averaging IP usage across the 500 replications.

The efficiency of the CAT was evaluated in all conditions by comparing the mean, minimum, and maximum number of items administered (NIA) over the 500 replications. Smaller mean values indicate more efficient tests. Typically, test efficiency is not evaluated with fixed length tests, whereas it is evaluated with variable length tests; however, the use of the forced answer (FA) item completion conditions in the study will result in variability in the NIA in both variable length and fixed length test conditions. Therefore, the efficiency of fixed length tests is also evaluated. Nonetheless, the test efficiency of fixed length tests will be of interest

mainly in forced answer (FA) conditions. The conditional NIA is also of interest because the abilities of examinees in the center of the item pool distribution will be measured better due to a larger number of items that match those ability levels. Conversely, examinees with extreme abilities, either very high or very low, will have fewer items in the item pool that match their ability levels. Therefore, the grand mean NIA, averaged over the 500 replications and conditioned on θ with 0.5 increments across the range of θ was plotted to assess conditional efficiency.

Chapter 4: Results

Nonconvergent Cases

Cases were considered nonconvergent if MLE was not implemented or the theta estimate was greater than +4 or less than -4 (i.e., out of range). Nonconvergent case were listwise deleted for all conditions before the outcome measures were calculated. For each condition, the mean number of nonconvergent cases across the 500 replications, for both types of nonconvergent cases, as well as the minimum and maximum number of nonconvergent cases are reported. Table 3 displays the mean number of nonconvergent MLE cases and the out-of-range cases averaged across the 500 replications in each condition. Across both types of nonconvergence cases, the grand mean of nonconvergent cases in conditions with IP sizes of 0, or the traditional CAT without implementing the IP method, was 49.92. The conditions in which IP size was 2 and 3 resulted in a grand means of 69.87 and 54.21 nonconvergent cases, respectively. On average, IP conditions with item pocket sizes of 4 resulted in fewer average nonconvergent cases, with a grand mean of 46.06. Across all conditions, the average number of cases where MLE was not reached was less than 1 across replications.

Averaging across conditions, the IP size 0 resulted in grand means of nonconvergent cases of 50.07 for the fixed length conditions, 49.77 for the variable length conditions, 49.65 for the 15 maximum items conditions, and 50.20 for the 20 maximum item conditions. The IP size of 2 resulted in an increase in nonconvergent cases. Across conditions, the IP size of 2 resulted in grand means of nonconvergent cases of 69.76 for the fixed length conditions, 69.97 for the variable length conditions, 70.36 for the 15 maximum items conditions, and 69.38 for the 20 maximum item conditions. There was a decrease in the grand means of nonconvergent cases

with the IP size of 3, resulting in grand means of 54.29 for the fixed length conditions, 54.13 for the variable length conditions, 54.50 for the 15 maximum items conditions, and 53.92 for the 20 maximum item conditions. The IP size of 4 resulted in the lowest grand means of nonconvergent cases, with 46.07 for the fixed length conditions, 46.06 for the variable length conditions, 46.51 for the 15 maximum items conditions, and 45.61 for the 20 maximum item conditions.

Using an item pocket size of 0, or the traditional CAT without the IP method, with the fixed length 15 item test resulted in a mean of 49.34 out-of-range cases, with a minimum of 33 and a maximum of 73. The fixed length tests with a maximum of 20 items resulted in a mean of 50.80 out-of-range cases, with a minimum of 34 and maximum of 74. The variable length maximum of 15 items tests resulted in a mean of 49.95 out-of-range cases across the 500 replications, and a minimum of 31 and a maximum of 71. The variable length maximum of 20 items resulted in a mean of 49.59 out-of-range cases, with a minimum of 30 and a maximum of 73 cases. For all traditional conditions, the mean number of nonconvergent MLE cases was less than 1 across the 500 replications (See Table 3).

The implementation of the Item Pocket method generally resulted in slightly more nonconvergent cases. The IP size of 2 with the 15 item fixed length test in the forced answer (FA) condition resulted in a mean of 70.34 out-of-range cases (min = 46, max = 93). The IP size of 2 with the 15 item fixed length test in the ignore (Ign) condition resulted in a slightly higher mean of 70.34 out-of-range cases (min = 48, max = 96). The IP size 2 with the 20 item fixed length test in the FA condition resulted in a slightly lower mean number of out-of-range cases of 69.17 (min = 51, max = 101). When ignoring the items in the pocket, the IP size of 2 with the 20 item fixed length test resulted in a mean of 69.26 out-of-range cases (min = 42, max = 94).

Similar patterns were displayed in the variable length conditions with IP sizes of 2, with a slightly higher mean of out-of-range cases for the 15 item test than the 20 item test. Specifically, the IP size of 2 with the 15 item fixed length test resulted in a mean of 70.35 out-of-range cases (min = 47, max = 100) when examinees were forced to answer items in the pocket whereas the mean number of out-of-range cases was 69.30 (min = 43, max = 95) for the corresponding condition with 20 items. In the same conditions (IP sizes of 2 with variable length tests), a slightly larger mean number of out-of-range cases resulted when examinees ignored the items in the pocket (*M* = 70.46, min = 46, max = 94) with 15 item tests as compared to 20 item tests (*M* = 69.77, min = 49, max = 93).

Overall, the IP size of 3 conditions resulted in lower mean out-of-range cases compared to the IP size of 2 conditions. For instance, in the forced answer condition with fixed length tests, the mean number of out-of-range cases was 53.96 (min = 35, max = 77) for the 15 item test while it was 54.21 (min = 36, max = 76) with the 20 item test. In the ignore IP condition with fixed length tests, the mean number of out-of-range cases was 54.82 (min = 30, max = 80) with the 15 item test whereas it was 54.18 (min = 36, max = 77) with the 20 item test. The mean number of out-of-range cases in the two variable length FA conditions were similar to those found in the two fixed length FA conditions, resulting in an average number of out-of-range cases of 54.31 (min = 36, max = 76) and 53.56 (min =34, max = 79) with 15 item and 20 item tests, respectively. The mean number of out-of-range cases in the variable length Ign conditions was 54.92 (min = 34, max = 79) and 53.74 (min =31, max = 75) with 15 item and 20 item tests, respectively.

Overall, the IP size of 4 conditions resulted in the lowest average number of out-of-range cases. The fixed length 15 item test conditions resulted in an average number of out-of-range

cases of 46.25 (min = 28, max = 67) and 46.75 (min = 27, max =66) for the FA and Ign conditions, respectively. The fixed length 20 item tests resulted in a slightly lower mean number of out-of-range cases as compared to the respective 15 item tests, with means of 45.22 (min = 28, max =64) and 46.05 (min = 23, max = 78) for the FA and Ign conditions, respectively. The variable length tests with 15 items resulted in similar numbers out-of-range cases as those found with the fixed length tests with 15 items. For instance, the mean number of out-of-range cases in the variable length FA test with 15 items condition was 46.65 (min = 26, max = 65). The variable length Ign test with 15 items condition resulted in a mean number of out-of-range cases equal to 46.40 (min = 28, max = 68). The variable length test conditions with 20 items resulted in an average number of out-of-range cases equal to 45.17 (min =28, max = 64) and 46.01 (min = 28, max = 68) in FA and Ign conditions, respectively.

The out-of-range cases resulting in the current study indicate that issues with ability estimation are present. Nonconvergence can result from the examinee responding in the extreme categories and, therefore, MLE is never implemented or the estimated thetas are out of range, meaning that the ability estimates are above $\theta = 4$ or below $\theta = -4$. The overall the mean number of nonconvergent cases across the 500 replications is approximately 5% of each condition. It should be noted that, on average, as the IP size increased, the mean number of nonconvergent cases decreased, indicating an interaction between the implementation of the IP method and nonconvergence. The issue of nonconvergent cases will be discussed further in the following chapter.

Condition			Out of Range			Nonconvergent MLE		
			Mean	Min	Max	Mean	Min	Max
Traditional (IP=0)	Fixed 15 Items		49.34	33	73	0.074	0	2
	Fixed 20 Items		50.80	34	74	0.034	0	1
	Variable 15 Items		49.95	31	71	0.068	0	2
	Variable 20 Items		49.59	30	73	0.034	0	2
IP Size 2	Fixed 15 Items	Forced Answer	70.28	46	93	0.062	0	1
		Ignored	70.34	48	96	0.106	0	3
	Fixed 20 Items	Forced Answer	69.17	51	101	0.032	0	1
		Ignored	69.26	42	94	0.034	0	1
	Variable 15 Items	Forced Answer	70.35	47	100	0.052	0	2
		Ignored	70.46	46	94	0.088	0	1
	Variable 20 Items	Forced Answer	69.30	43	95	0.028	0	1
		Ignored	69.77	49	93	0.028	0	1
IP Size 3	Fixed 15 Items	Forced Answer	53.96	35	77	0.048	0	2
		Ignored	54.82	30	80	0.080	0	2
	Fixed 20 Items	Forced Answer	54.21	36	76	0.032	0	1
		Ignored	54.18	34	77	0.040	0	1
	Variable 15 Items	Forced Answer	54.31	36	76	0.056	0	1
		Ignored	54.92	34	79	0.080	0	2
	Variable 20 Items	Forced Answer	53.56	34	79	0.026	0	1
		Ignored	53.74	31	75	0.034	0	2
IP Size 4	Fixed 15 Items	Forced Answer	46.25	28	67	0.050	0	2
		Ignored	46.75	27	66	0.088	0	3
	Fixed 20 Items	Forced Answer	45.23	28	64	0.030	0	1
		Ignored	46.05	23	78	0.042	0	2
	Variable 15 Items	Forced Answer	46.65	26	65	0.044	0	2
		Ignored	46.40	28	68	0.084	0	2
	Variable 20 Items	Forced Answer	45.17	28	64	0.028	0	1
		Ignored	46.01	28	68	0.032	0	2

Table 3. Nonconvergent Cases Averaged Across the 500 Replications

Estimated Thetas

The overall recovery of the known thetas was evaluated with descriptive statistics. The grand mean, across the 500 replications, as well as the average standard deviations of the estimated thetas is presented in Table 4. The mean standard error of the theta estimates within each condition is included in Table 4, in addition to the minimum and maximum theta estimates and standard errors across the 500 replications. All conditions resulted in slightly larger grand mean theta estimates and standard deviations compared to the known theta grand mean of 0.0 and standard deviation of 1.0.

Overall, the known theta estimates were recovered slightly better in the IP size of 0 (traditional CAT) conditions than in the other IP size conditions, with grand means closer to zero and lower standard deviations. The fixed length conditions with an IP size of 0 resulted in grand mean theta estimates of 0.017 and 0.014 with corresponding standard deviations of 1.096 and 1.080, for the maximum items of 15 and 20, respectively. The variable length conditions, with a maximum of 15 and 20 items, resulted in grand mean theta estimates of 0.015 (*SD* = 1.095) and 0.009 (*SD* = 1.077), respectively. The traditional fixed length and variable length 20 item conditions resulted in the lowest grand means, 0.014 and 0.009, respectively, as expected due to increased measurement precision as the number of items administered increases. Conversely, the shorter tests, both fixed length 15 items and variable length 15 items, resulted in the largest grand mean theta estimates.

The implementation of the IP resulted in a slight increase in grand means compared to the traditional CATs without the implementation of the IP method. The IP size of 2 resulted in grand mean theta estimates of 0.018 (*SD* = 1.095) and 0.020 (*SD* = 1.106) for the fixed length 15 items forced answer (FA) and ignore (Ign) conditions, respectively. The IP size 2 fixed length

20 items conditions resulted in grand mean theta estimates of 0.016 (*SD* = 1.083) and 0.018 (*SD* = 1.089) for the FA and Ign conditions, respectively. The same pattern was seen with the variable length conditions, with larger grand means and standard deviations resulting in the 15 item tests than the 20 item tests. The variable length conditions with 15 maximum items resulted in grand mean theta estimates of 0.018 (*SD* = 1.095) and 0.019 (*SD* = 1.105) for the FA and Ign conditions, respectively. The variable length 20 item test resulted in slightly lower grand mean theta estimates of 0.011 (*SD* = 1.081) and 0.012 (*SD* = 1.088) for the FA and Ign conditions, respectively. Of the IP size 2 conditions, the variable length with a maximum of 20 items resulted in the lowest grand mean theta estimates and standard deviations.

As IP size increased, the grand mean of the theta estimates decreased, approaching the grand means of the traditional, baseline conditions. The IP size of 3, fixed length 15 item test resulted in grand mean theta estimates of 0.017 (*SD* = 1.086) and 0.019 (*SD* = 1.101) for the FA and Ign conditions, respectively. The IP size 3 fixed length test with 20 items resulted in identical grand mean theta estimates for the FA and Ign conditions of 0.015 with slightly different standard deviations, FA (*SD* = 1.075) and Ign (*SD* = 1.083). Again, the variable length conditions with an IP size of 3 resulted in a similar pattern as was seen with the IP size of 2, with the variable length conditions performing slightly better than the fixed length conditions. The variable length test with 15 maximum items resulted in grand mean theta estimates of 0.016 (*SD* = 1.086) and 0.019 (*SD* = 1.101) for the FA and Ign conditions, respectively. In contrast, the variable length conditions with a maximum of 20 items resulted in grand mean theta estimates of 0.011 (*SD* = 1.073) and 0.011 (*SD* = 1.083) for the FA and Ign conditions, respectively. The variable length test with 20 maximum items resulted in the lowest grand mean theta estimates and standard deviations for the IP size of 3 conditions.

Similar to the IP size of 2 and 3 conditions, the IP size of 4 generally resulted in slightly lower grand mean theta estimates and standard deviations with the longer test conditions (20 maximum items) and the variable length conditions. For the IP size of 4, fixed length with 15 item tests, the grand mean theta estimates were 0.016 (*SD* = 1.080) and 0.019 (*SD* = 1.099) for the FA and Ign conditions, respectively. The fixed length 20 item test conditions resulted in grand mean theta estimates of 0.014 (*SD* = 1.070) and 0.016 (*SD* = 1.082) for the FA and Ign conditions, respectively. The variable length conditions with an IP size of 4 and maximum of 15 items resulted in grand mean theta estimates of 0.015 (*SD* = 1.080) and 0.018 (*SD* = 1.099) for the FA and Ign conditions, respectively. The variable length 20 maximum item tests with an IP size of 4 resulted in grand mean theta estimates of 0.011 (*SD* = 1.068) and 0.012 (*SD* = 1.080) for the FA and Ign conditions, respectively.

The recovery of the known theta's for the Ignore conditions resulted in slightly larger theta estimate grand means and standard deviations as compared to the FA conditions. For instance, the IP size of 2 with fixed length 15 item test Ign condition resulted in a theta estimate grand mean of 0.020 (*SD* = 1.106), whereas the corresponding FA condition resulted in a grand mean of 0.018 (*SD* = 1.095). This same pattern is repeated for all IP size conditions, with very slight decreases in grand means and standard deviations as IP size increases.

The overall precision of measurement was assessed with the mean standard error, averaged over the 500 replications within each condition (see Table 4). In addition, the minimum and maximum for each condition across the 500 replications illustrates the range of standard errors across replications. The traditional, IP size of 0, fixed length 15 item test condition resulted in a mean standard error of 0.333 (min = 0.268, max = 0.895) while the fixed length 20 item tests resulted in a mean standard error of 0.295 (min = 0.244, max = 0.760). The

traditional, IP size of 0, variable length 15 maximum item test condition resulted in a slightly higher mean standard error of 0.338 (min = 0.281, max = 0.859). The variable length with 20 maximum item test conditions resulted in a mean standard error of 0.317 (min = 0.278, max = 0.761), slightly lower than the shorter variable length condition, but slightly higher than the fixed length condition with 20 maximum items.

Again, as IP sizes increased, the standard errors generally decreased. Specifically, the IP size of 2 with fixed length 15 item test conditions resulted in a mean standard error of 0.315 (min = 0.257,max = 0.810) and 0.338 (min = 0.272, max = 0.857) for the FA and Ign conditions, respectively. The IP size of 2 for the fixed length 20 item test conditions resulted in mean standard errors of 0.284 (min = 0.236, max = 0.735) and 0.300 (min = 0.247, max = 0.760) for the FA and Ign conditions, respectively. The variable length 15 maximum item test conditions resulted in mean standard errors of 0.318 (min = 0.254, max = 0.815) and 0.342 (min = 0.283, max = 0.853) for the FA and Ign conditions, respectively, which is a slight increase compared to the fixed length 15 item test conditions. The IP size of 2 variable length with 20 maximum item tests resulted in mean standard errors of 0.299 (min = 0.252, max = 0.733) and 0.319 (min = 0.280, max = 0.758) for the FA and Ign conditions, respectively.

The IP size of 3 resulted in slightly lower standard errors than those seen with IP size of 2, however, this is only the case with the FA conditions. The IP size of 3 fixed length 15 item test conditions resulted in mean standard errors of 0.307 (min = 0.252, max = 0.788) and 0.341 (min = 0.276, max = 0.852) for the FA and Ign conditions, respectively. A slight decrease is seen with the fixed length 20 item test conditions, with mean standard errors of 0.279 (min = 0.233, max = 0.720) and 0.302 (min = 0.251, max = 0.751), respectively, for the FA and Ign conditions. The variable length conditions in conjunction with the IP size of 3 resulted in

slightly increased standard errors, on average, compared to the fixed length conditions. For the variable length 15 maximum item tests, the mean standard errors were 0.308 (min = 0.248, max = 0.787) and 0.344 (min = 0.286, max = 0.858) for the FA and Ign conditions, respectively. The IP size of 3 variable length 20 maximum item tests resulted in slightly lower mean standard errors of 0.291 (min = 0.246, max = 0.721) and 0.319 (min = 0.283, max = 0.756) for the FA and Ign conditions, respectively.

The IP size of 4 with fixed length 15 item test conditions resulted in a slightly lower mean standard error of 0.300 (min = 0.247, max = 0.770) and 0.344 (min = 0.280, max = 0.858) for the FA and Ign conditions, respectively. The fixed length 20 maximum item tests with an IP size of 4, resulted in mean standard errors of 0.274 (min = 0.230, max = 0.709) and 0.305 (min = 0.254, max = 0.763), respectively, for the FA and Ign conditions. The variable length conditions with the IP size of 4 produced similar mean standard errors as the fixed length conditions. Specifically, the variable length 15 maximum item test conditions resulted in mean standard errors of 0.300 (min = 0.245, max = 0.765) and 0.345 (min = 0.288, max = 0.860) for the FA and Ign conditions, respectively. The variable length 20 maximum item test conditions resulted in mean standard errors of 0.284 (min = 0.240, max = 0.707) and 0.320 (min = 0.285, max = 0.763), respectively, for the FA and Ign conditions.

Similar to the traditional CAT conditions, increasing the maximum number of items resulted in more precise measurement as indicated by smaller standard errors. The fixed length 15 item FA test conditions resulted in mean standard errors of 0.315, 0.307, and 0.300 for IP sizes of 2, 3,and 4, respectively. The fixed length 20 item FA tests resulted in the most precise measurement for all IP size conditions, with mean standard errors of 0.284, 0.279, and 0.274 for IP sizes of 2, 3, and 4 respectively. This same pattern is seen with the variable length conditions,

with decreases in SEs as test length increases. For instance, the variable length 15 maximum item FA test conditions resulted in mean SEs of 0.318, 0.308, and 0.300, respectively, for IP sizes of 2, 3, and 4. The mean standard errors for IP sizes of 2, 3, and 4 for the variable length 20 maximum item FA test conditions were 0.299, 0.291, and 0.284, respectively.

However, the opposite is also true. That is, as the maximum number of items decreased, the IP size increased, and the items in the pocket are ignored, the standard errors increased, demonstrating a slight loss in measurement precision. The fixed length 15 item Ign test conditions resulted in mean standard errors of 0.338, 0.341, and 0.344 for the IP sizes of 2, 3, and 4, respectively. Increasing test length decreased the standard errors; however, under the Ign conditions, as IP size increased, the SEs increases slightly. For instance, the fixed length 20 item Ign test conditions resulted in mean standard errors of 0.300, 0.302, and 0.305, respectively, for IP sizes of 2, 3, and 4. The variable length 15 item Ign test conditions resulted in the largest mean standard errors for all IP size conditions, with means of 0.342, 0.344, and 0.345 for IP sizes of 2, 3, and 4, respectively. The same general pattern is seen with the variable length 20 maximum item tests under the Ign condition; however, the impact on the standard errors is diminished, with mean SEs of 0.319, 0.319 and 0.320 for IP sizes of 2, 3, and 4, respectively.

For the FA conditions, as the IP sizes increased, the SEs slightly decreased. Conversely, the fixed length 20 item Ign tests resulted in a mean SE of 0.300, 0.302, and 0.305 for IP sizes of 2, 3, and 4, respectively. The implementation of the IP method under the FA conditions resulted in increased measurement precision, as seen with lower mean standard errors, which is due to the additional information gained with the forced administration of additional items. Conversely, the implementation of the IP method under the Ignore conditions resulted in a slight overall loss in

measurement precision when compared to the traditional CAT without an IP, which is due to the loss of information from the items that were ultimately ignored.

Condition			Final θ Estimate			Standard Error		
			Grand Mean (SD)	Min	Max	Mean	Min	Max
Traditional (IP=0)	Fixed 15 Items		0.017(1.096)	-3.540	3.662	0.333	0.268	0.859
	Fixed 20 Items		0.014(1.080)	-3.457	3.607	0.295	0.244	0.760
	Variable 15 Items		0.015(1.095)	-3.540	3.667	0.338	0.281	0.859
	Variable 20 Items		0.009(1.077)	-3.449	3.609	0.317	0.278	0.761
IP Size 2	Fixed 15 Items	Forced Answer	0.018(1.095)	-3.489	3.629	0.315	0.257	0.810
		Ignored	0.020(1.106)	-3.530	3.667	0.338	0.272	0.857
	Fixed 20 Items	Forced Answer	0.016(1.083)	-3.468	3.593	0.284	0.236	0.735
		Ignored	0.018(1.089)	-3.457	3.609	0.300	0.247	0.760
	Variable 15 Items	Forced Answer	0.018(1.095)	-3.500	3.647	0.318	0.254	0.815
		Ignored	0.019(1.105)	-3.547	3.657	0.342	0.283	0.853
	Variable 20 Items	Forced Answer	0.011(1.081)	-3.465	3.589	0.299	0.252	0.733
		Ignored	0.012(1.088)	-3.456	3.601	0.319	0.280	0.758
IP Size 3	Fixed 15 Items	Forced Answer	0.017(1.086)	-3.479	3.610	0.307	0.252	0.788
		Ignored	0.019(1.101)	-3.546	3.654	0.341	0.276	0.852
	Fixed 20 Items	Forced Answer	0.015(1.075)	-3.452	3.580	0.279	0.233	0.720
		Ignored	0.015(1.083)	-3.460	3.568	0.302	0.251	0.751
	Variable 15 Items	Forced Answer	0.016(1.086)	-3.466	3.605	0.308	0.248	0.787
		Ignored	0.019(1.101)	-3.546	3.675	0.344	0.286	0.858
	Variable 20 Items	Forced Answer	0.011(1.073)	-3.460	3.582	0.291	0.246	0.721
		Ignored	0.011(1.083)	-3.469	3.586	0.319	0.283	0.756
IP Size 4	Fixed 15 Items	Forced Answer	0.016(1.080)	-3.461	3.599	0.300	0.247	0.770
		Ignored	0.019(1.099)	-3.558	3.666	0.344	0.280	0.858
	Fixed 20 Items	Forced Answer	0.014(1.070)	-3.435	3.573	0.274	0.230	0.709
		Ignored	0.016(1.082)	-3.477	3.601	0.305	0.254	0.763
	Variable 15 Items	Forced Answer	0.015(1.080)	-3.474	3.584	0.300	0.245	0.765
		Ignored	0.018(1.099)	-3.535	3.671	0.345	0.288	0.860
	Variable 20 Items	Forced Answer	0.011(1.068)	-3.438	3.563	0.284	0.240	0.707
		Ignored	0.012(1.080)	-3.452	3.601	0.320	0.285	0.763

Table 4. Grand Mean Theta Estimates and Standard Error Descriptive Statistics Averaged Across the 500 Replications

Overall Measurement Precision

The Pearson product-moment correlations between the known and estimated thetas are presented in Table 5. The mean correlations, across the 500 replications, illustrate the accuracy in recovering the known thetas. In addition to the mean correlation, the minimum and maximum correlation across the 500 replications are reported in Table 5. The traditional CAT, with an IP size of 0, fixed length condition resulted in a mean correlation of 0.946 (min = 0.921, max = 0.957) for the 15 item test and 0.957 (min = 0.934, max = 0.967) for the 20 item test. The IP size of 0 variable length conditions resulted in slightly lower mean correlations of 0.945 (min = 0.896, max = 0.956) for the 15 maximum item test condition and 0.951 (min = 0.909, max = 0.961) for the 20 maximum item test condition.

Implementation of the IP method generally increased accuracy in recovering the known thetas. The IP size of 2 fixed length 15 item test conditions resulted in mean correlations of 0.951 (min = 0.904, max = 0.961) and 0.944 (min = 0.907, max = 0.958) for the FA and Ign conditions, respectively. The fixed length 20 item test conditions resulted in slightly increased mean correlations of 0.959 (min = 0.928, max = 0.969) and 0.954 (min = 0.917, max = 0.965), respectively, for the FA and Ign conditions. The IP size of 2 variable length conditions resulted in a slight decrease in mean correlations compared to their fixed length counterparts, with mean correlations of 0.949 (min = 0.914, max = 0.961) and 0.943 (min = 0.914, max = 0.956) for the variable length 15 maximum item FA and Ign test conditions, respectively. The IP size of 2 variable length 20 maximum item test conditions resulted in mean correlations of 0.955 (min = 0.920, max = 0.965) and 0.950 (min = 0.920, max = 0.961) for the FA and Ign conditions, respectively. Overall, the known thetas were more accurately recovered in the Forced Answer IP test conditions, resulting in slightly higher correlations as compared to the traditional CAT

conditions. Conversely, slightly lower correlations were found in the Ignore conditions when compared to the traditional CAT conditions; however, the correlations were comparable across the IP sizes. The same pattern continued in the IP size of 3 Forced Answer conditions, which resulted in higher mean correlations as compared to the IP sizes of 0 and 2 forced answer conditions. Specifically, the IP size of 3 fixed length 15 items conditions resulted in mean correlations of 0.954 (min = 0.924, max = 0.963) and 0.944 (min = 0.913, max = 0.955) for the FA and Ign conditions, respectively. The mean correlations increased slightly for the fixed length 20 item test conditions, resulting in mean correlations of 0.961 (min = 0.925, max = 0.969) and 0.955 (min = 0.925, max = 0.964) for the FA and Ign conditions, respectively. Again, a slight decrease in mean correlations is seen in the variable conditions compared to the fixed length conditions. The IP size of 3 variable length 15 maximum item tests resulted in mean correlations of 0.953 (min = 0.926, max = 0.963) for the FA condition and 0.943 (min = 0.917, max = 0.957) for the Ign condition.

Generally, the pattern with the IP sizes of 2 and 3 is also seen in IP size of 4 conditions. The IP size of 4 fixed length 15 item test conditions resulted in mean correlations of 0.956 (min = 0.934, max = 0.965) for the FA and 0.944 (min = 0.910, max = 0.955) for the Ign condition. The fixed length 20 item test conditions resulted in increased mean correlations, with a mean of 0.962 (min = 0.939, max = 0.970) and 0.954 (min = 0.925, max = 0.964) for the FA and Ign conditions, respectively. The IP size of 4 variable length 15 maximum item test conditions resulted in similar, but slightly lower correlations than the fixed length conditions with means of 0.955 (min = 0.929, max = 0.964) for the FA condition and 0.943 (min = 0.913, max = 0.955) for the Ign condition. The variable length 20 maximum item test conditions resulted in slightly larger mean correlations than the shorter variable length conditions, with mean correlations of

0.960 (min = 0.925, max = 0.967) and 0.950 (min = 0.925, max = 0.960) for the FA and Ign conditions, respectively.

The impact of the item completion conditions on the mean correlation is more distinctive with the FA conditions than the Ign conditions. For instance, the fixed length 15 item FA test conditions resulted in mean correlations of 0.951, 0.954, and 0.956 for IP sizes 2, 3, and 4, respectively. The mean correlation between the known and estimated thetas increased as IP size increased. This pattern continues as test length increased, with mean correlations of 0.959, 0.961, and 0.962 for fixed length 20 item FA test conditions with IP sizes of 2, 3, and 4, respectively. The variable length 15 maximum item FA test conditions resulted in mean correlations of 0.949, 0.953, and 0.955, respectively, for IP sizes of 2, 3, and 4. Again, the mean correlations increased as test length increased, with mean correlations of 0.955, 0.957, and 0.960 for the variable length 20 maximum item FA test conditions for IP sizes of 2, 3, and 4 respectively. Conversely, the fixed length 15 item Ign test conditions resulted in mean correlations of 0.944 for IP sizes 2, 3, and 4. Increasing test length to 20 items for the fixed length test conditions under the Ign conditions resulted in mean correlations of 0.954, 0.955, and 0.954 for IP sizes of 2, 3, and 4, respectively. The Ign conditions for the variable length test conditions resulted in identical mean correlations for IP sizes of 2, 3, and 4 equal to 0.943 for the 15 maximum item test conditions and 0.950 for the 20 maximum item test conditions, displaying no impact on recovery of known thetas for the variable length conditions under the Ign conditions.

Condition			Correlation		
			Mean	Min	Max
Traditional (IP=0)	Fixed 15 Items		0.946	0.921	0.957
	Fixed 20 Items		0.957	0.934	0.967
	Variable 15 Items		0.945	0.896	0.956
	Variable 20 Items		0.951	0.909	0.961
IP Size 2	Fixed 15 Items	Forced Answer	0.951	0.904	0.961
		Ignored	0.944	0.907	0.958
	Fixed 20 Items	Forced Answer	0.959	0.928	0.969
		Ignored	0.954	0.917	0.965
	Variable 15 Items	Forced Answer	0.949	0.914	0.961
		Ignored	0.943	0.914	0.956
	Variable 20 Items	Forced Answer	0.955	0.920	0.965
		Ignored	0.950	0.910	0.961
IP Size 3	Fixed 15 Items	Forced Answer	0.954	0.924	0.963
		Ignored	0.944	0.913	0.955
	Fixed 20 Items	Forced Answer	0.961	0.925	0.969
		Ignored	0.955	0.925	0.964
	Variable 15 Items	Forced Answer	0.953	0.926	0.963
		Ignored	0.943	0.917	0.957
	Variable 20 Items	Forced Answer	0.957	0.925	0.967
		Ignored	0.950	0.914	0.961
IP Size 4	Fixed 15 Items	Forced Answer	0.956	0.934	0.965
		Ignored	0.944	0.910	0.955
	Fixed 20 Items	Forced Answer	0.962	0.939	0.970
		Ignored	0.954	0.925	0.964
	Variable 15 Items	Forced Answer	0.955	0.929	0.964
		Ignored	0.943	0.913	0.955
	Variable 20 Items	Forced Answer	0.960	0.925	0.967
		Ignored	0.950	0.925	0.960

Table 5. Pearson product-moment Correlations between Known and Estimated Thetas Averaged Across 500 Replications

The average, minimum, and maximum Bias and RMSE associated with the final theta estimates were calculated across the 500 replications for each condition and are presented in Table 6. The mean bias for the IP size of 0 fixed length 15 item test condition was -0.015, whereas the fixed length 20 item test condition resulted in a mean bias of -0.012. The same pattern is seen with the variable length conditions with an IP size of 0. Specifically, mean bias for the maximum of 15 item test condition was -0.013 and was -0,007 for the maximum 20 item test condition.

The IP size of 2 fixed length 15 item test conditions resulted in mean biases of -0.013 and -0.016 for the FA and Ign conditions, respectively. The IP size of 2 fixed length 20 item test conditions resulted in a mean bias of -0.012 for both FA and Ign conditions. The IP size of 2 variable length 15 maximum item test conditions resulted in similar mean bias as the fixed length conditions, with a mean bias of -0.012 for the FA and a mean bias of -0.015 for the Ign condition. The mean bias decreased slightly for the IP size of 2 variable length 20 maximum item test conditions, with mean bias of -0.007 and -0.008 for the FA and Ign conditions, respectively.

The IP size of 3 conditions produced similar mean bias as that of IP size of 2 conditions. Specifically, the IP size of 3 fixed length 15 item test resulted in mean bias of -0.012 for the FA condition and mean bias of -0.016 for the Ign condition. The IP size of 3 fixed length 20 item test conditions resulted in mean bias of -0.010 and -0.012 for the FA and Ign conditions, respectively. The IP size of 3 variable length 15 maximum item test conditions resulted in mean bias of -0.012 and -0.015 for the FA and Ign conditions, respectively. The variable length 20 maximum item test with an IP size of 3 resulted in a mean bias of -0.007 for both the FA and Ign conditions.

The IP size of 4 fixed length 15 item test conditions resulted in a mean bias of -0.012 for the FA condition and a mean bias of -0.015 for the Ign condition. The mean bias for the IP size of 4 fixed length 20 item test conditions was -0.010 and -0.012 for the FA and Ign conditions, respectively. The IP size of 4 variable length 15 maximum item test conditions resulted in a mean bias of -0.011 for the FA condition and a mean bias of -0.015 for the Ign condition. The IP size of 4 variable length 20 maximum item test conditions resulted in mean bias of -0.007 and -0.008 for the FA and Ign conditions, respectively.

The impact to mean bias under the two item completion conditions is minimal. Under the FA conditions, the mean bias for the fixed length 15 item test was -0.013, -0.012, and -0.012 for IP sizes of 2, 3, and 4, respectively. Increasing test length reduced mean bias very slightly, with the fixed length 20 item FA test conditions resulting in mean bias of -0.012, -0.010, and -0.010 respectively, for IP sizes of 2, 3, and 4. The variable length conditions resulted in similar mean bias across IP size conditions under the FA conditions. For instance, the variable length 15 maximum item FA test conditions resulted in mean bias of -0.012, -0.012, and -0.011 for IP sizes 2, 3, and 4, respectively. In contrast, increasing test length to 20 maximum items resulted in a mean bias of -0.007 for all IP size conditions under the FA condition. Under the Ign condition, the decrease in mean bias seen in the FA conditions is less apparent. For instance, the fixed length 15 item Ign test conditions resulted in mean bias of -0.016, -0.016, and -0.015 for IP sizes of 2, 3, and 4, respectively. Increasing test length to 20 items for the fixed length tests resulted in mean bias of -0.012 for all IP size conditions. The variable length 15 maximum item Ign test conditions resulted in mean bias of -0.015 for all IP size conditions as well. The variable length 20 maximum item Ign test conditions resulted in mean bias of -0.008, -0.007, and -0.008 for IP sizes of 2, 3, and 4, respectively, which is a slight decrease with an increase in test length.

The mean RMSE displayed a consistent pattern (see Table 6), with the Forced Answer conditions resulting in lower mean RMSEs when compared to the traditional CATs and the Ignore conditions resulting in mean RMSEs slightly higher than those in the traditional CAT conditions. The traditional CAT (with IP size of 0) conditions resulted in a mean RMSE of 0.356 for the fixed length 15 item test and a mean RMSE of 0.314 for the fixed length 20 item test. The variable length 15 maximum item test condition resulted in a mean RMSE of 0.358 and a mean RMSE of 0.332 resulted for the variable length 20 maximum item test condition. The pattern seen here with the fixed length conditions resulting in slightly lower mean RMSE than the variable length conditions, as well as the longer test conditions resulting in lower RMSE than the shorter test conditions, is seen in all IP size conditions.

The IP size of 2 conditions displayed the same pattern seen with the traditional conditions with the longer fixed length test conditions resulting in a lower mean RMSE, on average. The IP size of 2 fixed length 15 item test conditions resulted in mean RMSEs of 0.340 and 0.365 for the FA and Ign conditions, respectively. Increasing the test length to 20 items resulted in a mean RMSE of 0.308 for the FA condition and a mean RMSE of 0.326 for the Ign condition. The variable length conditions resulted in slightly higher mean RMSEs than the fixed length conditions with a mean RMSE of 0.345 for the variable length 15 item FA test and a mean RMSE of 0.368 for the variable length 15 item Ign test condition. Increasing test length decreased mean RMSE to 0.319 for the variable length 20 maximum item FA test condition and to 0.340 for the variable length 20 maximum item Ign test condition; however, the impact on the mean RMSE is not as substantial for the variable length conditions as it is in the fixed length conditions.

The above pattern continued with the IP size of 3 conditions, with the fixed length 15 item tests resulting in mean RMSEs of 0.327 and 0.364 for the FA and Ign conditions, respectively. The fixed length 20 item test conditions resulted in a mean RMSE of 0.297 for the FA condition and a mean RMSE of 0.323 for the Ign condition. The IP size of 3 with variable length test conditions resulted in a mean RMSE of 0.329 for the 15 maximum item FA tests and a mean RMSE of 0.367 for the 15 maximum item Ign tests, whereas the 20 maximum item tests resulted in a mean RMSE of 0.309 and 0.337 for the FA and Ign conditions, respectively. Again, the longer fixed length test condition resulted in lower mean RMSE compared to the longer variable length test conditions.

In addition to the pattern of the longer fixed length conditions resulting in lower mean RMSE, the RMSE also decreased as IP size increased. For instance, the IP size of 4 fixed length 15 item test conditions resulted in a mean RMSE of 0.319 and 0.345 for the FA and Ign conditions, respectively. The fixed length 20 item test conditions resulted in a lower mean RMSE of 0.293 for the FA condition and a mean RMSE of 0.323 for the Ign condition. The variable length 15 maximum item test resulted in a mean RMSE of 0.321 for the FA condition and a mean RMSE of 0.366 for the Ign condition, whereas the variable length 20 maximum item test conditions resulted in mean RMSEs of 0.301 and 0.336 for the FA and Ign conditions, respectively.

The two item completion conditions had an impact on the resulting mean RMSE. For instance, the fixed length 15 item FA test conditions resulted in mean RMSEs of 0.340, 0.327, and 0.319 for IP sizes of 2, 3, and 4, respectively, decreasing with IP size increases. Conversely, the fixed length 15 item Ign test conditions resulted in mean RMSEs of 0.365, 0.364, and 0.345 for IP sizes of 2, 3, and 4, respectively, displaying a more gradual decrease with IP size

increases. The fixed length 20 item FA test conditions resulted in mean RMSEs of 0.308, 0.297,and 0.293 for IP sizes of 2, 3, and 4, respectively. The fixed length 20 item Ign test conditions resulted in mean RMSE of 0.326, 0.323, and 0.323, respectively, for IP sizes 2, 3, and 4. The variable length 15 maximum item FA test conditions resulted in RMSEs of 0.345, 0.329, and 0.321 for IP sizes 2, 3, and 4, respectively. The Ign conditions for the variable length 15 maximum item tests resulted in mean RMSEs of 0.368, 0.367, and 0.366 for IP sizes of 2, 3, and 4, respectively. The variable length 20 maximum item FA tests resulted mean RMSEs of 0.319, 0.309, and 0.301 for IP sizes of 2, 3, and 4, respectively. In contrast, the Ign conditions for the variable length 20 maximum item tests resulted in a slower decrease in mean RMSE as IP size increased, with mean RMSEs of 0.340, 0.337, and 0.336 for IP sizes 2, 3, and 4, respectively. Again, as IP size increased, the mean RMSE decreased, in both the Forced Answer and Ignore item completions conditions; however, the decrease under the Ign conditions was more gradual.

Condition			Bias			RMSE		
			Mean	Min	Max	Mean	Min	Max
Traditional (IP=0)	Fixed 15 Items		-0.015	-2.357	1.267	0.356	0.319	0.434
	Fixed 20 Items		-0.012	-2.112	1.097	0.314	0.278	0.405
	Variable 15 Items		-0.013	-2.226	1.272	0.358	0.326	0.483
	Variable 20 Items		-0.007	-2.099	1.113	0.332	0.298	0.438
IP Size 2	Fixed 15 Items	Forced Answer	-0.013	-2.387	1.206	0.340	0.298	0.476
		Ignored	-0.016	-2.560	1.298	0.365	0.318	0.464
	Fixed 20 Items	Forced Answer	-0.012	-2.333	1.072	0.308	0.270	0.403
		Ignored	-0.012	-2.564	1.117	0.326	0.289	0.429
	Variable 15 Items	Forced Answer	-0.012	-2.569	1.235	0.345	0.308	0.460
		Ignored	-0.015	-2.591	1.285	0.368	0.329	0.450
	Variable 20 Items	Forced Answer	-0.007	-2.321	1.085	0.319	0.280	0.423
		Ignored	-0.008	-2.437	1.135	0.340	0.304	0.459
IP Size 3	Fixed 15 Items	Forced Answer	-0.012	-2.189	1.171	0.327	0.295	0.440
		Ignored	-0.016	-2.335	1.301	0.364	0.322	0.453
	Fixed 20 Items	Forced Answer	-0.010	-1.984	1.064	0.297	0.264	0.410
		Ignored	-0.012	-2.208	1.131	0.323	0.288	0.411
	Variable 15 Items	Forced Answer	-0.012	-2.167	1.160	0.329	0.295	0.404
		Ignored	-0.015	-2.399	1.301	0.367	0.325	0.447
	Variable 20 Items	Forced Answer	-0.007	-2.118	1.059	0.309	0.274	0.413
		Ignored	-0.007	-2.157	1.152	0.337	0.303	0.447
IP Size 4	Fixed 15 Items	Forced Answer	-0.012	-2.071	1.121	0.319	0.283	0.392
		Ignored	-0.015	-2.256	1.283	0.345	0.326	0.456
	Fixed 20 Items	Forced Answer	-0.010	-2.029	1.011	0.293	0.261	0.367
		Ignored	-0.012	-2.055	1.130	0.323	0.289	0.425
	Variable 15 Items	Forced Answer	-0.011	-2.074	1.138	0.321	0.293	0.400
		Ignored	-0.015	-2.220	1.308	0.366	0.323	0.439
	Variable 20 Items	Forced Answer	-0.007	-2.004	1.040	0.301	0.271	0.419
		Ignored	-0.008	-2.103	1.126	0.336	0.300	0.411

Table 6. Mean, Minimum, and Maximum Bias and RMSE Averaged Across 500 Replications

Conditional Measurement Precision

Conditional plots of mean bias (see Figures 8, 9, 10, and 11) and grand mean SE (see Figures 12, 13, 14, and 15) associated with the final theta estimates, averaged across the 500 replications, were created in order to examine the performance of the IP method at different ability levels. Plots of mean bias conditional on known theta for each of the four stopping rule conditions are shown in Figures 8, 9, 10, and 11. Figure 8A displays the mean bias for the fixed length 15 item FA and Ign test conditions for all IP sizes. As can be seen in the top figure of Figure 8A, in the FA conditions, the conditional mean bias across the range of theta (-3.0 to +3.0) is the same as in the traditional CAT conditions (with IP size of 0). For known thetas below $\theta = -3.0$ and above $\theta = +3.0$ in the Forced Answer conditions, the mean bias departed slightly from that of the conditional traditional CAT, with more positive bias for the higher abilities and slightly less negative bias for the lower abilities. In the Ignore conditions (see Figure 8A, bottom plot), the mean conditional bias generally mirrored that of the traditional CATs for most ability levels, with slight departures for IP size 3 with abilities less than $\theta = -2.5$ where slightly more negative bias was observed.

When comparing the Forced Answer and Ignore conditions for the fixed length 15 item tests (see Figure 8B), the Forced Answer conditions resulted in slightly less negative bias for ability levels ranging from 0 to 2.0 and slightly less positive bias for ability levels ranging from $\theta = -2.5$ to $\theta = 0$ with an IP size of 2. As IP size increased, the Ign conditions resulted in slightly more positive bias between $\theta = -2.5$ and $\theta = -1.0$ and slightly more negative bias for abilities between $\theta = 0.05$ and $\theta = 2.0$ as compared to the FA conditions (see Figure 8C & 8D). However, as the test length increased, these differences disappeared. As shown in Figure 9A, in the fixed length 20 item test conditions under both FA and Ign conditions, all IP sizes produced very

similar mean conditional bias across the range of theta from θ = -3.0 to θ = +3.0 . The extreme abilities displayed slight differences with the IP sizes of 2, 3, and 4, with the FA test conditions resulting in slightly less negative bias for abilities below θ = -3.0 and slightly more positive bias for abilities above θ = +3.0 as compared to the traditional condition. These differences declined under the Ign conditions. Comparing the FA and Ign conditions (see Figure 9B, 9C, and 9D), as IP size increased the differences at the extremes of the ability distribution decreased and the mean conditional bias in the center of the distribution are practically identical.

This same pattern is seen in the variable length test conditions with the 15 item stopping rule as was seen above with the fixed length test conditions. As seen in Figure 10A, the conditional bias is very similar for abilities in the center of the ability distribution; however, the extreme abilities result in slightly more negative bias in ability levels below θ = -3.0 and slightly more positive bias in abilities above θ = +3.0 for IP sizes 2, 3, and 4 under the FA conditions as compared to the traditional conditions. These differences disappear under the Ign conditions, with the IP sizes' lines practically overlapping (see Figure 10A, bottom plot). Comparing the mean conditional bias for the FA and Ign test conditions with IP sizes of 2, 3, and 4 (see Figures 10B, 10C, and 10D), as IP size increased, slightly more negative bias is seen in the higher ability levels (θ = 0 to θ = 2.5) and slightly more positive bias for the lower ability levels (θ = 0 to θ = -2.5) under the Ign conditions.

As test length increased, the differences seen between the Forced Answer and Ignore conditions with IP sizes of 2, 3, and 4 decreased (see Figure 11A). The differences in the center of the ability distribution for the variable length 15 maximum item test conditions practically disappears in the longer 20 maximum item test conditions. As seen in Figure 11A, the IP size conditions 2, 3, and 4 are overlapping the traditional condition for both the FA and Ign

conditions across the majority of the ability distribution. Slight departures are seen for the extreme ability levels. As seen in Figures 11B, 11C, and 11D, as IP size increased, the differences under the FA and Ign conditions in mean conditional bias for the extreme abilities decreased, with the FA conditions resulting in less mean conditional bias at the extreme ability levels.

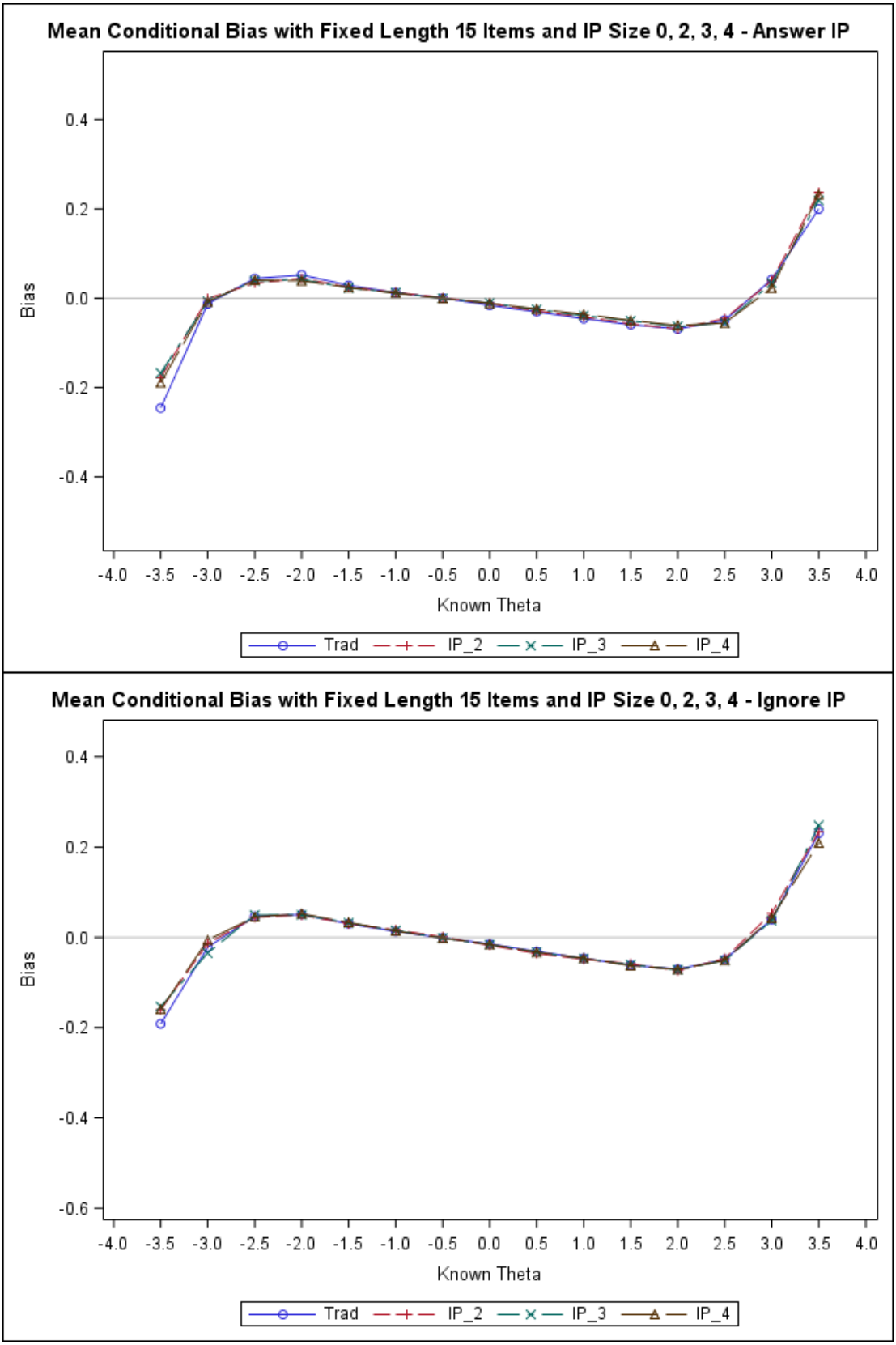

Figure 8A. Plots of Mean Bias Conditional on Known Theta for Fixed Length 15 Items, IP Size 0, 2, 3, & 4, Forced Answer & Ignore Conditions

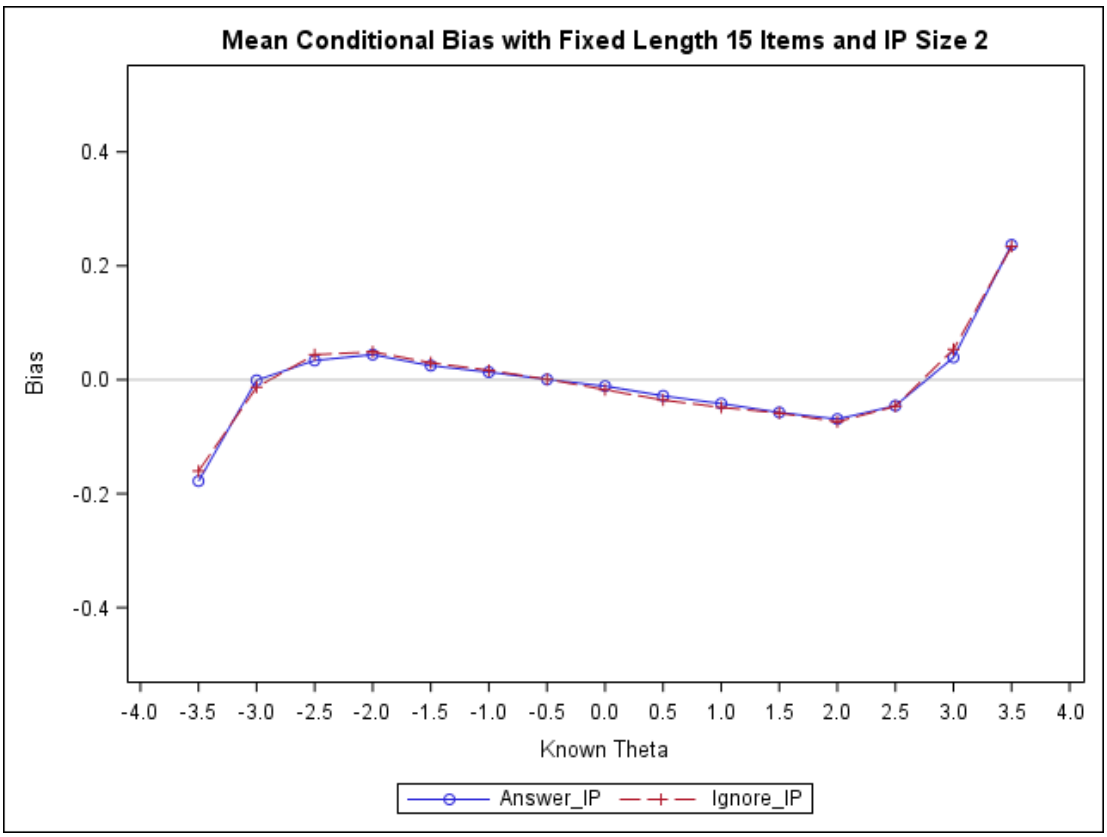

Figure 8B. Plot of Mean Bias Conditional on Known Theta for Fixed Length 15 Items, IP Size 2, Forced Answer & Ignore Conditions

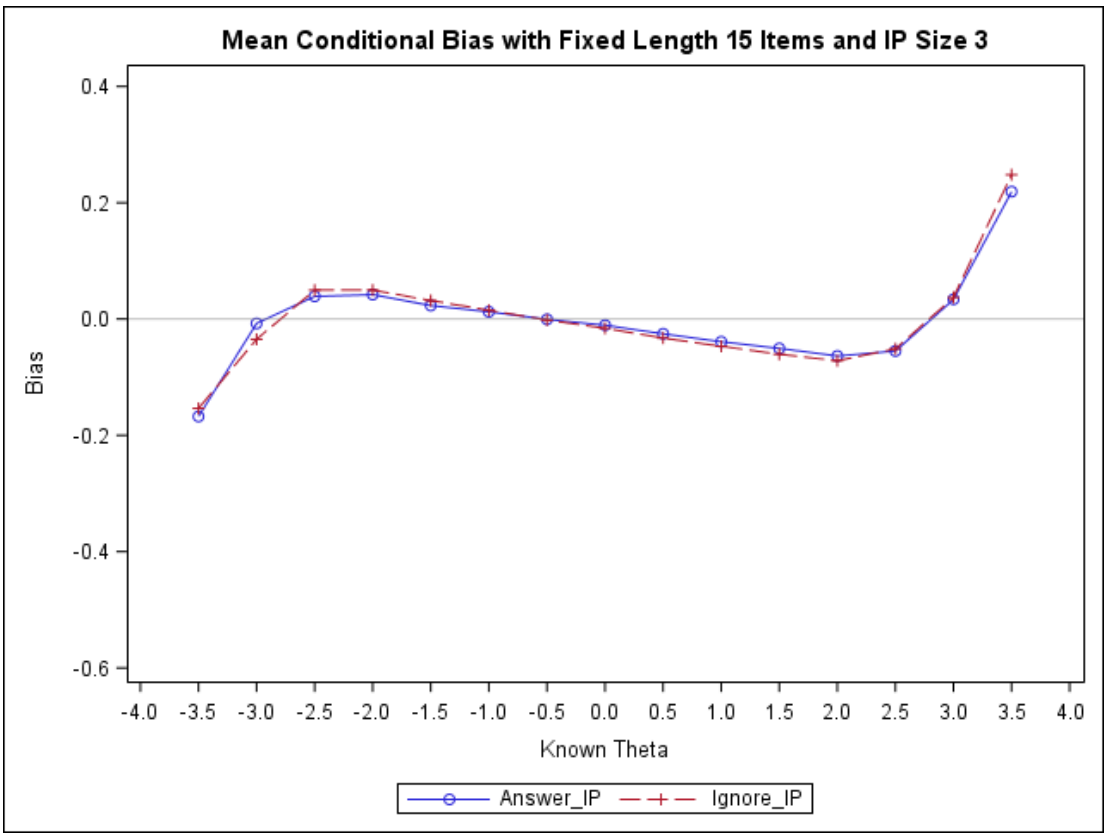

Figure 8C. Plot of Mean Bias Conditional on Known Theta for Fixed Length 15 Items, IP Size 3, Forced Answer & Ignore Conditions

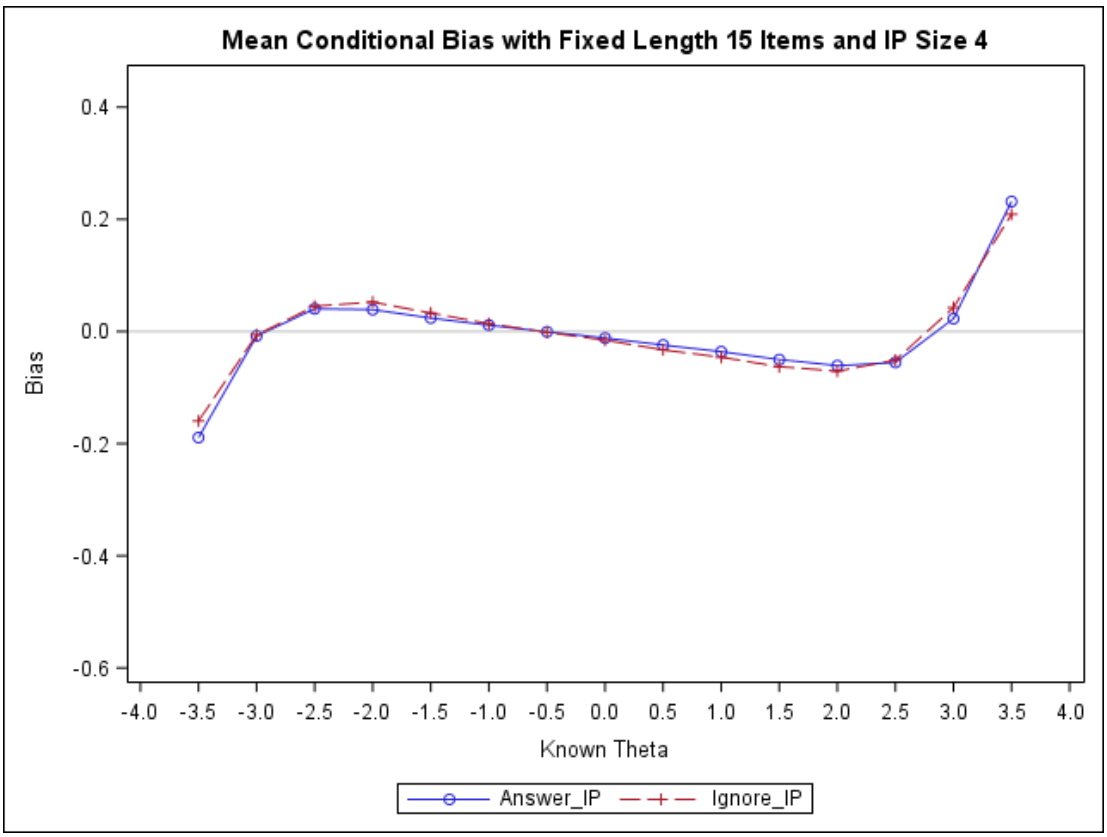

Figure 8D. Plot of Mean Bias Conditional on Known Theta for Fixed Length 15 Items, IP Size 4, Forced Answer & Ignore Conditions

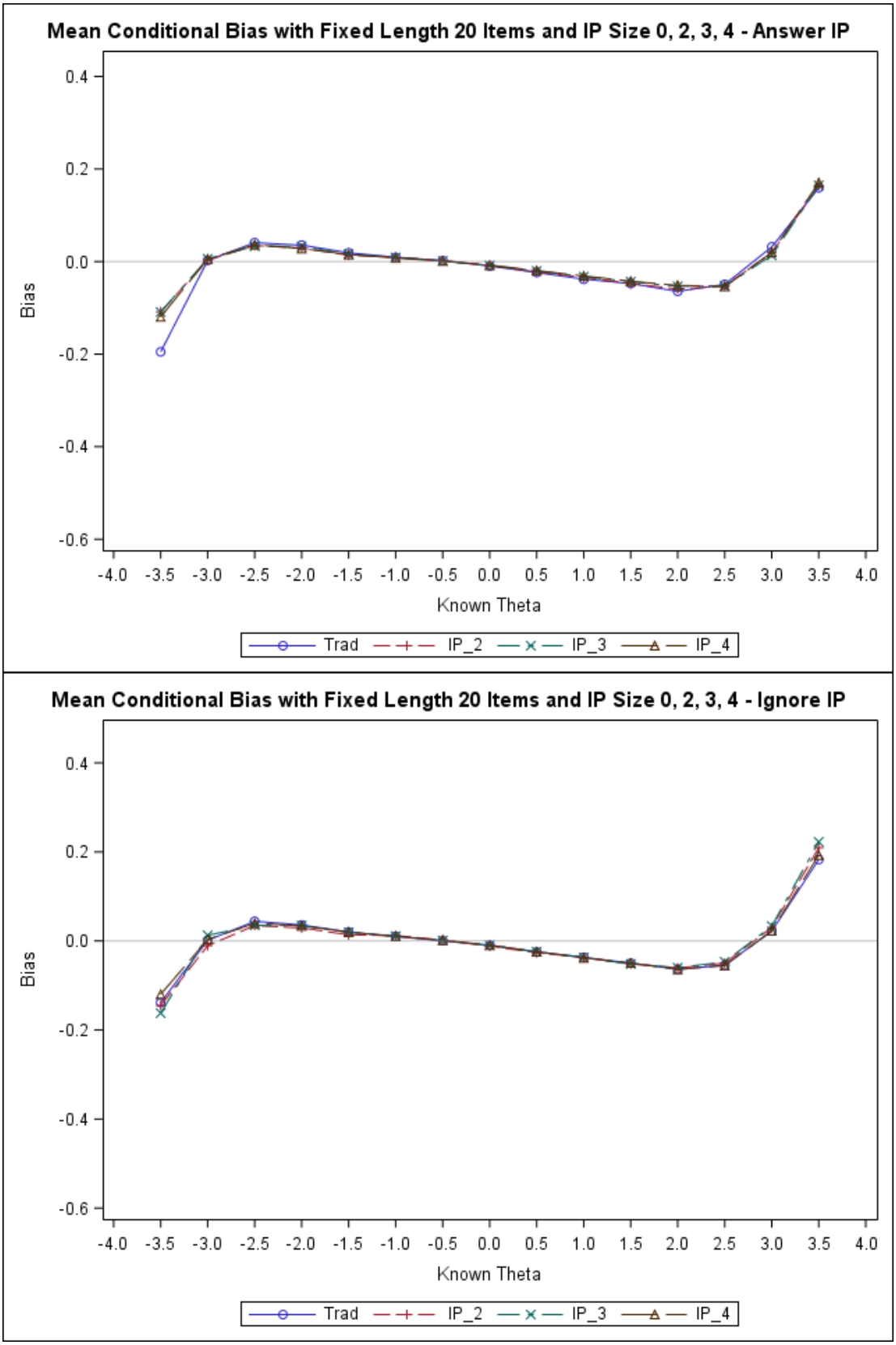

Figure 9A. Plots of Mean Bias Conditional on Known Theta for Fixed Length 20 Items, IP Size 0, 2, 3, & 4, Forced Answer & Ignore Conditions

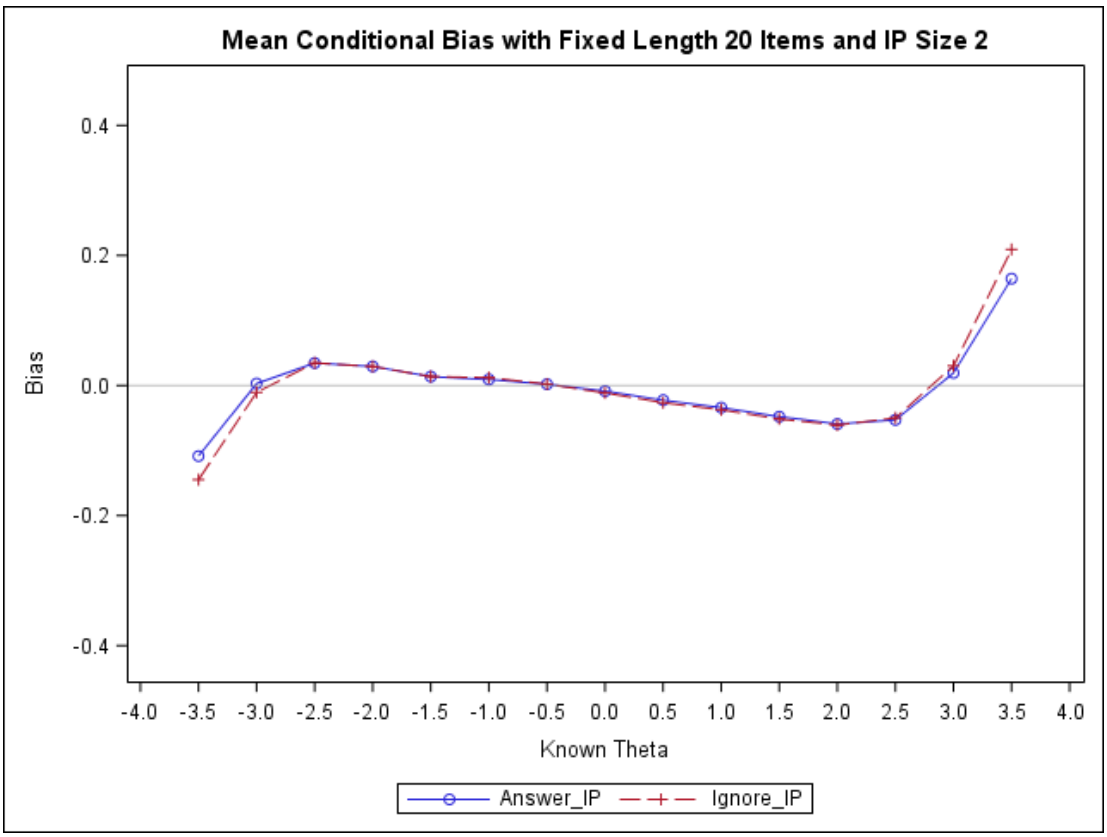

Figure 9B. Plot of Mean Bias Conditional on Known Theta for Fixed Length 20 Items, IP Size 2, Forced Answer & Ignore Conditions

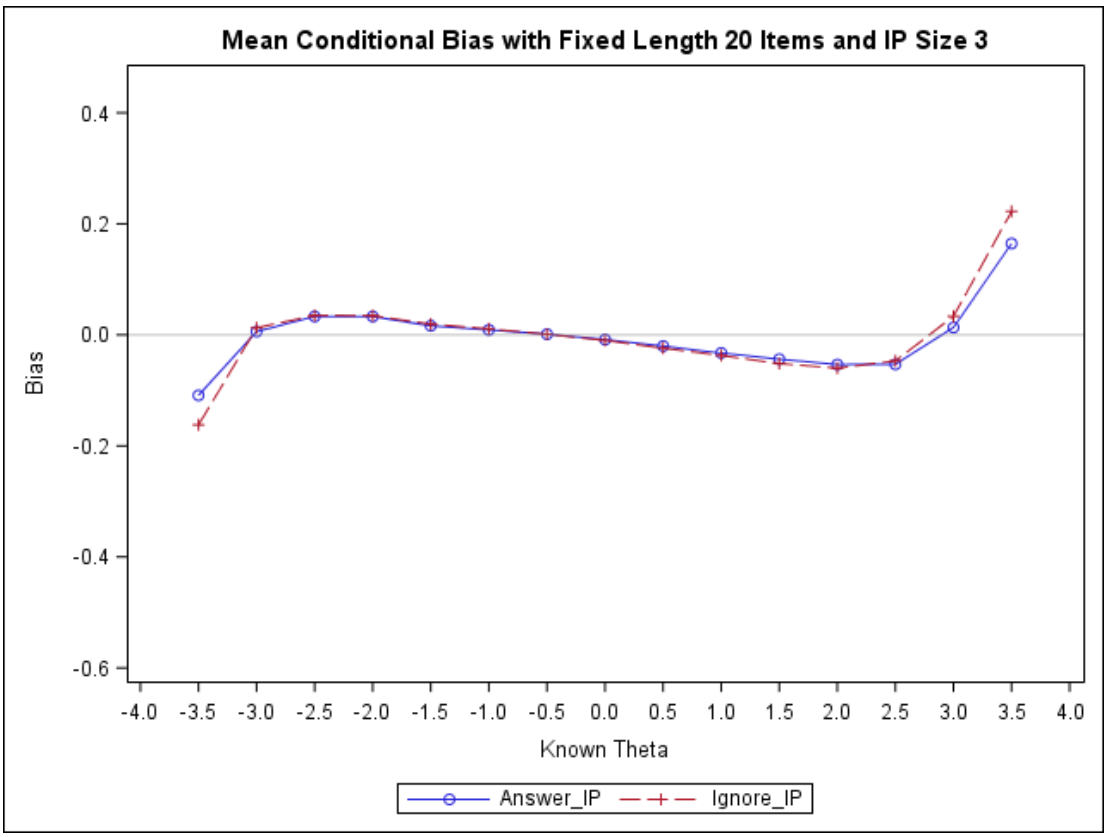

Figure 9C. Plot of Mean Bias Conditional on Known Theta for Fixed Length 20 Items, IP Size 3, Forced Answer & Ignore Conditions

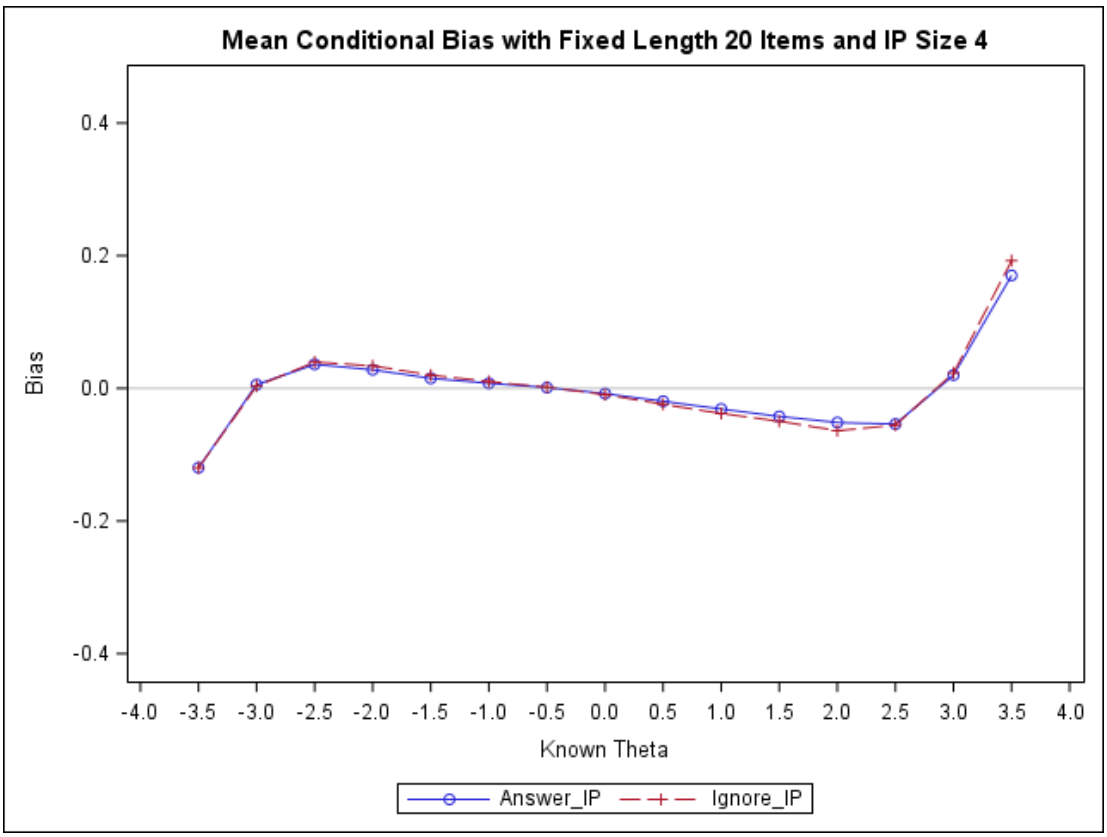

Figure 9D. Plot of Mean Bias Conditional on Known Theta for Fixed Length 20 Items, IP Size 4, Forced Answer & Ignore Conditions

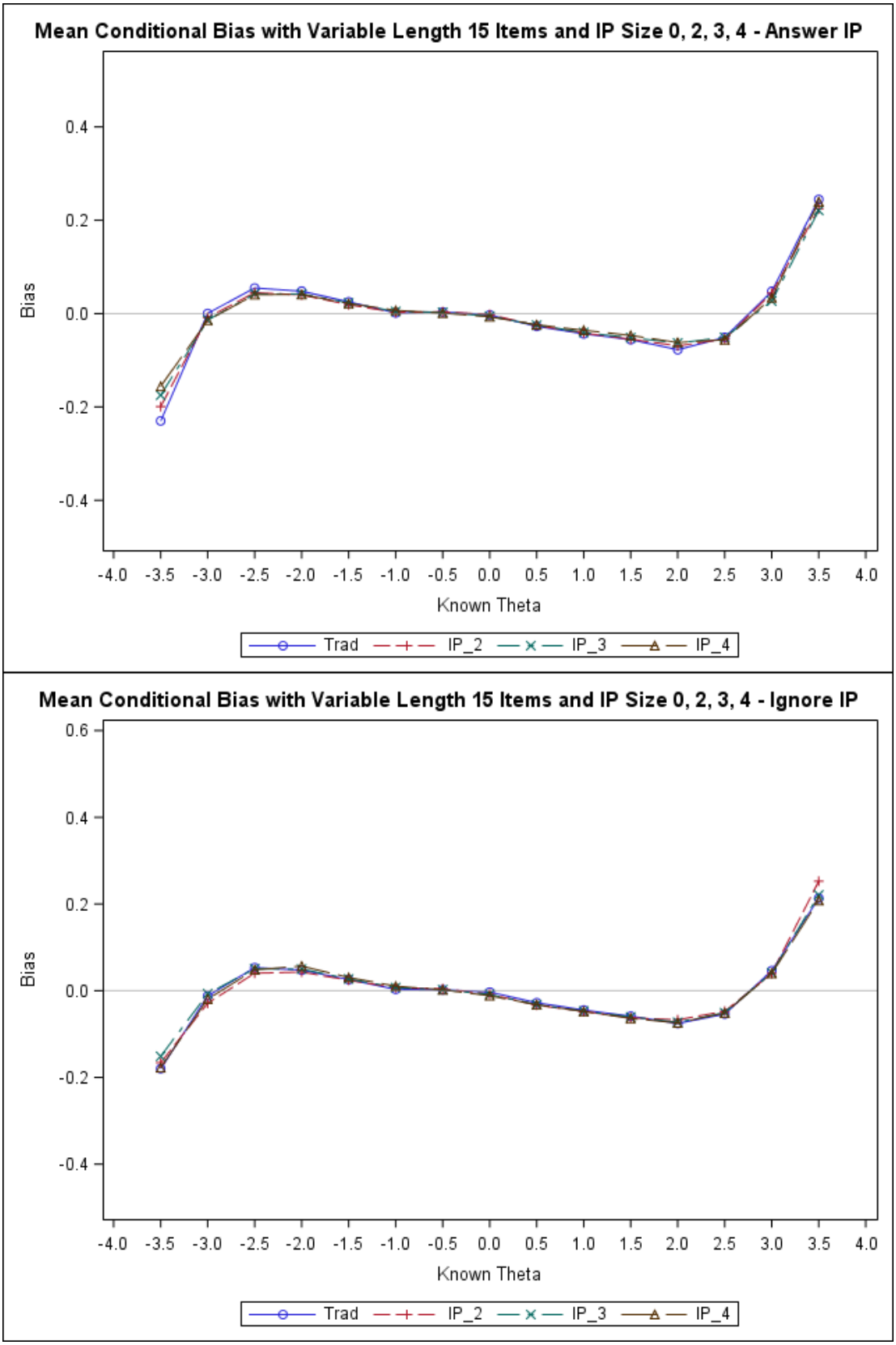

Figure 10A. Plots of Mean Bias Conditional on Known Theta for Variable Length 15 Items, IP Size 0, 2, 3, & 4, Forced Answer & Ignore Conditions

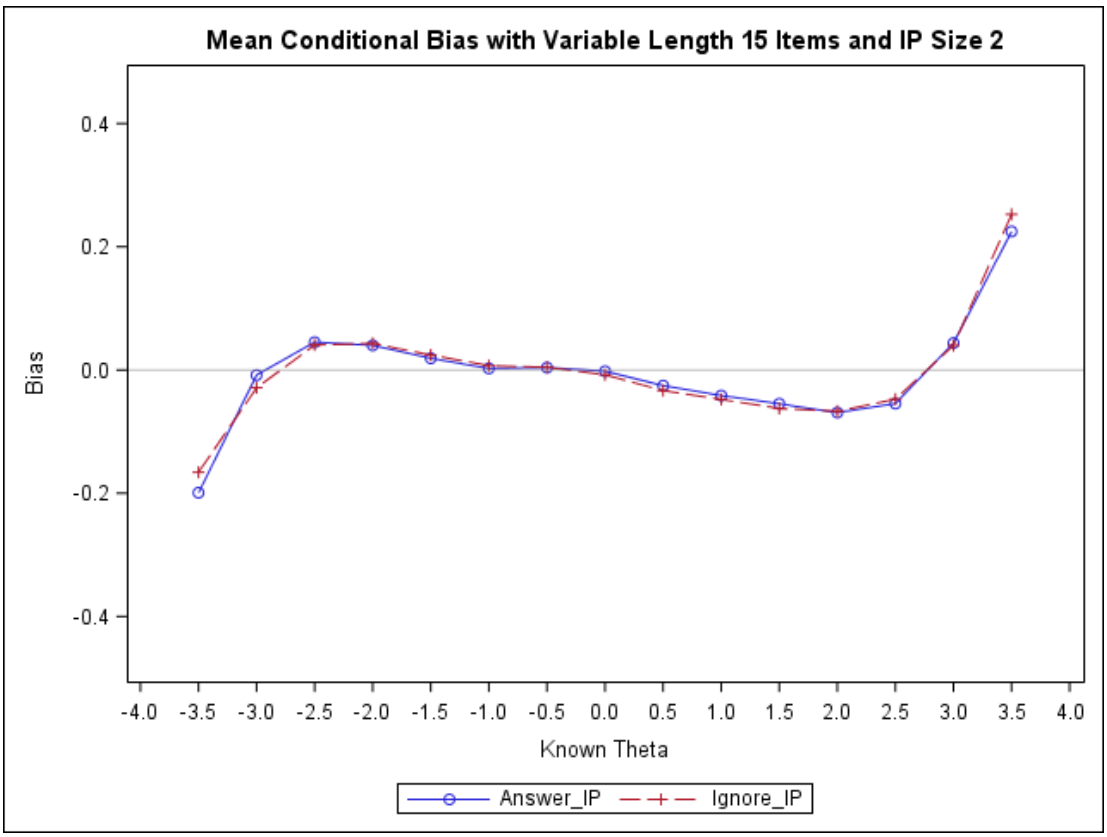

Figure 10B. Plot of Mean Bias Conditional on Known Theta for Variable Length 15 Items, IP Size 2, Forced Answer & Ignore Conditions

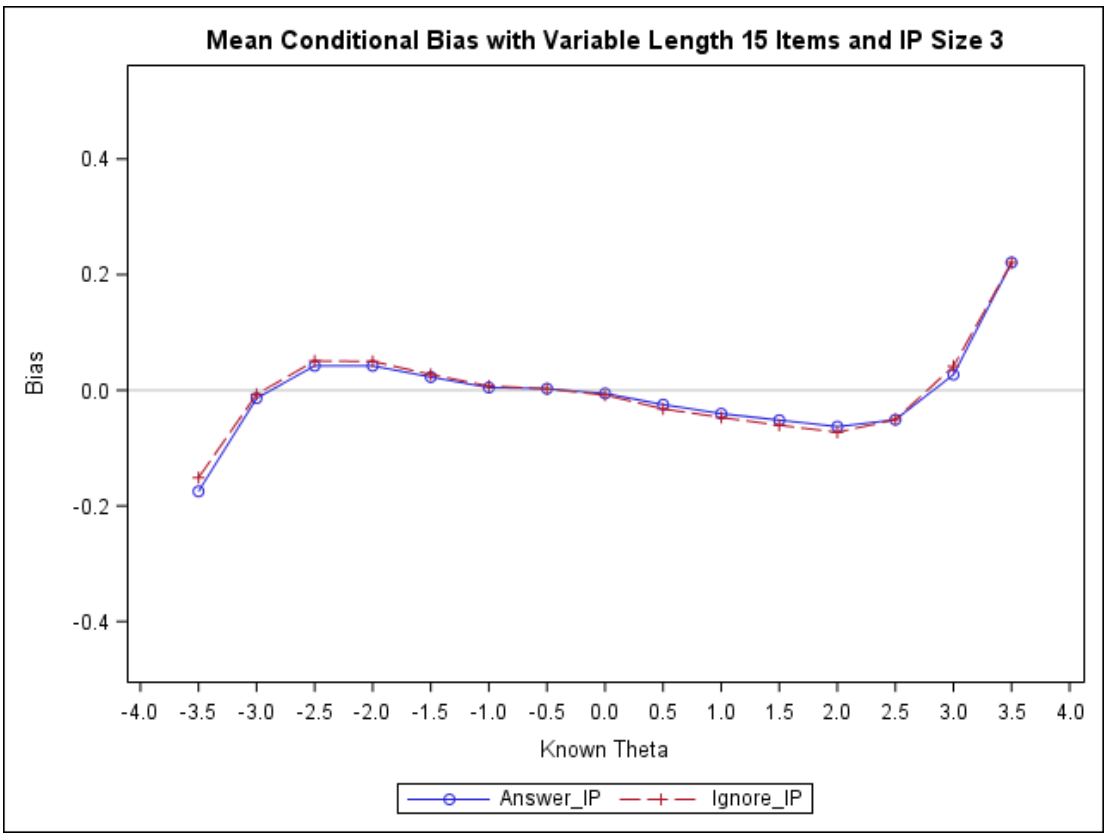

Figure 10C. Plot of Mean Bias Conditional on Known Theta for Variable Length 15 Items, IP Size 3, Forced Answer & Ignore Conditions

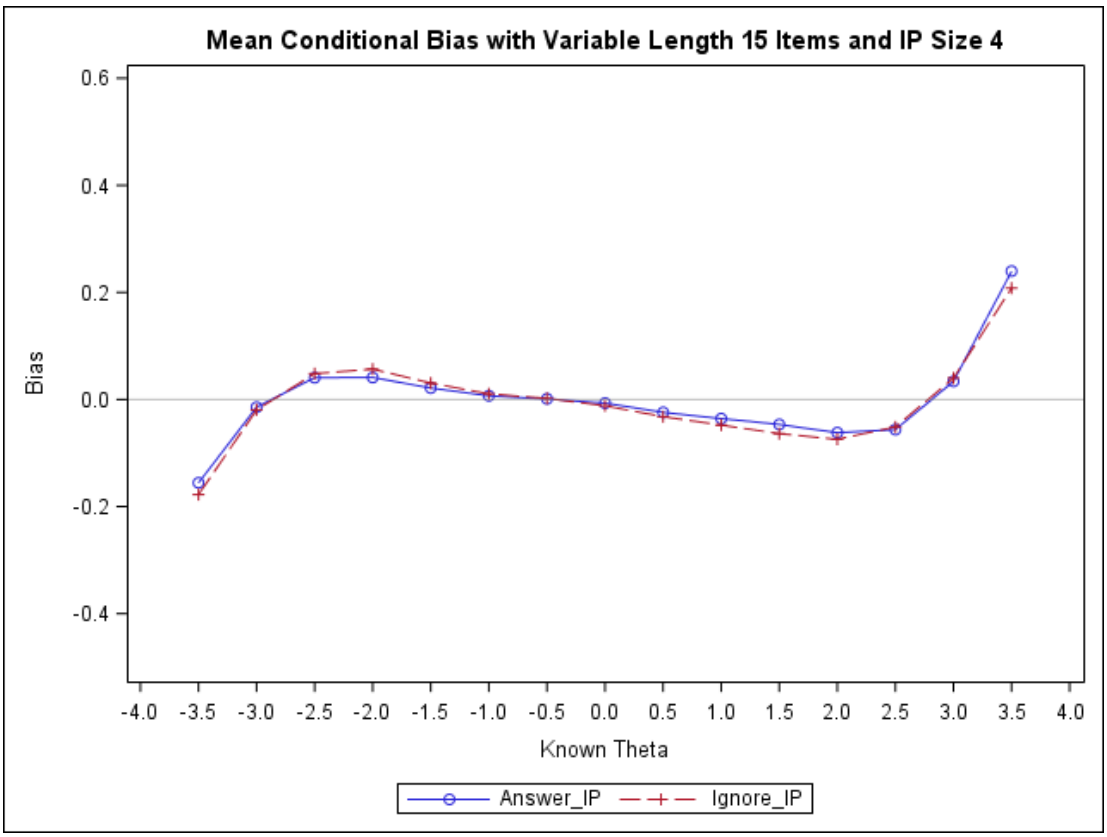

Figure 10D. Plot of Mean Bias Conditional on Known Theta for Variable Length 15 Items, IP Size 4, Forced Answer & Ignore Conditions

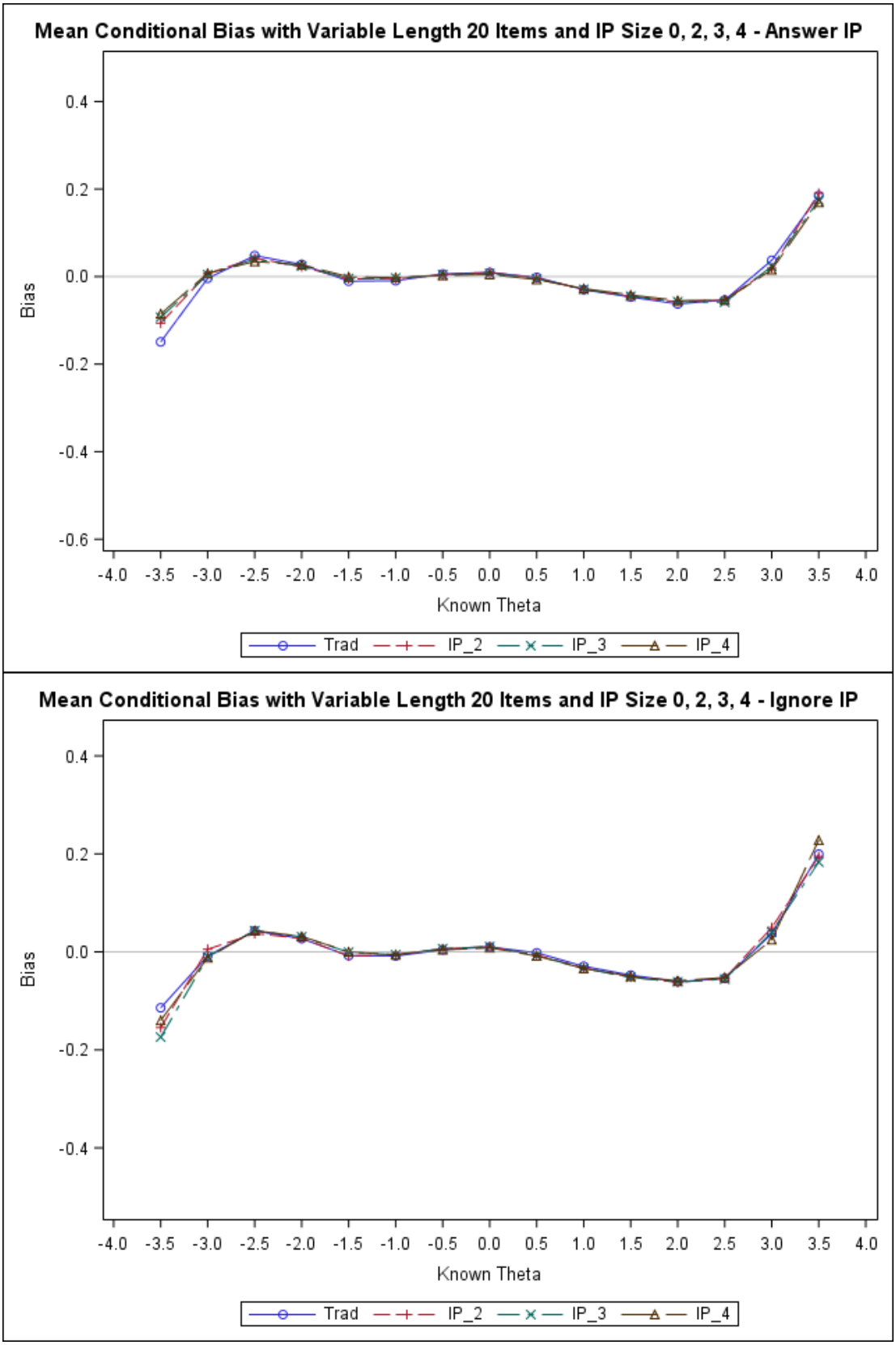

Figure 11A. Plots of Mean Bias Conditional on Known Theta for Variable Length 20 Items, IP Size 0, 2, 3, & 4, Forced Answer & Ignore Conditions

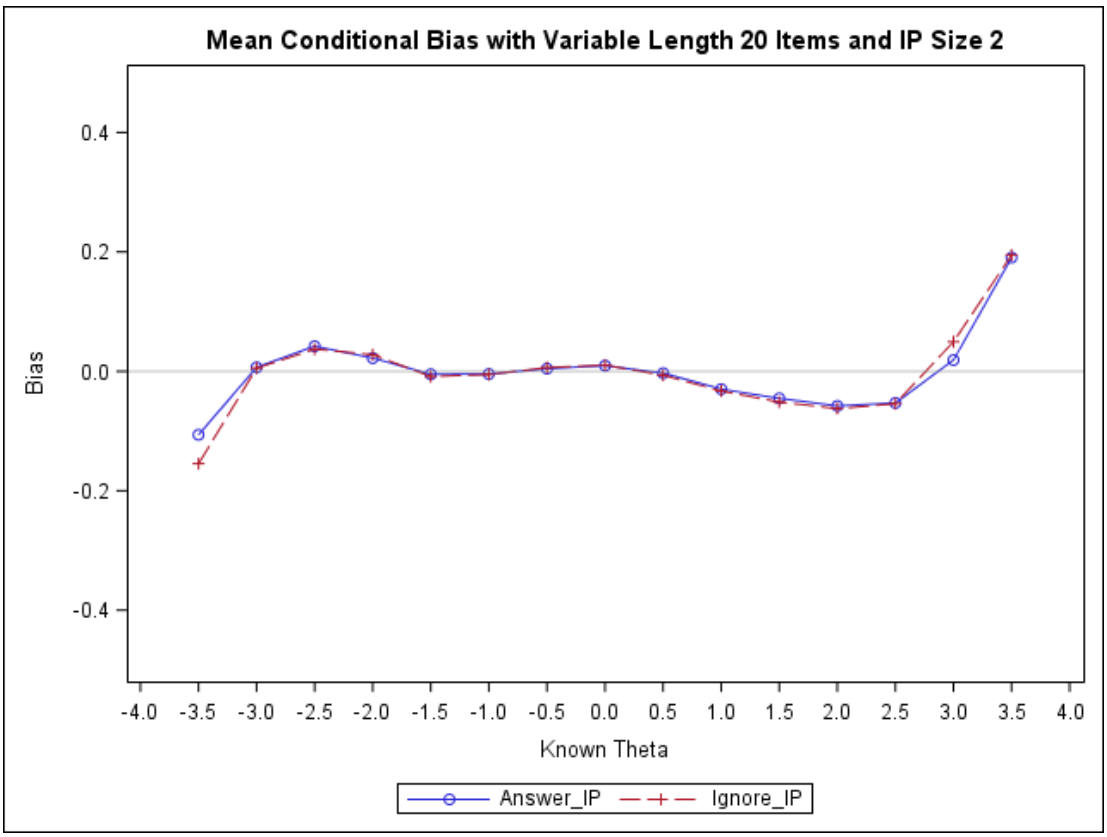

Figure 11B. Plot of Mean Bias Conditional on Known Theta for Variable Length 20 Items, IP Size 2, Forced Answer & Ignore Conditions

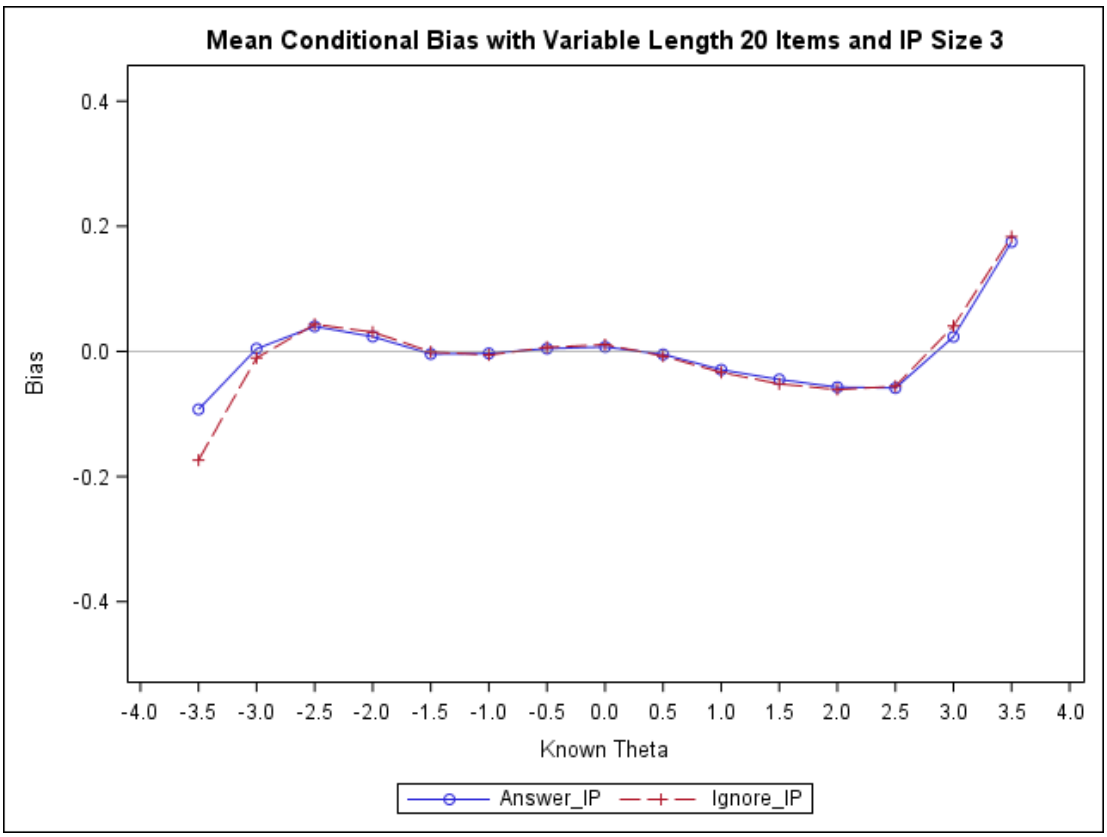

Figure 11C. Plot of Mean Bias Conditional on Known Theta for Variable Length 20 Items, IP Size 3, Forced Answer & Ignore Conditions

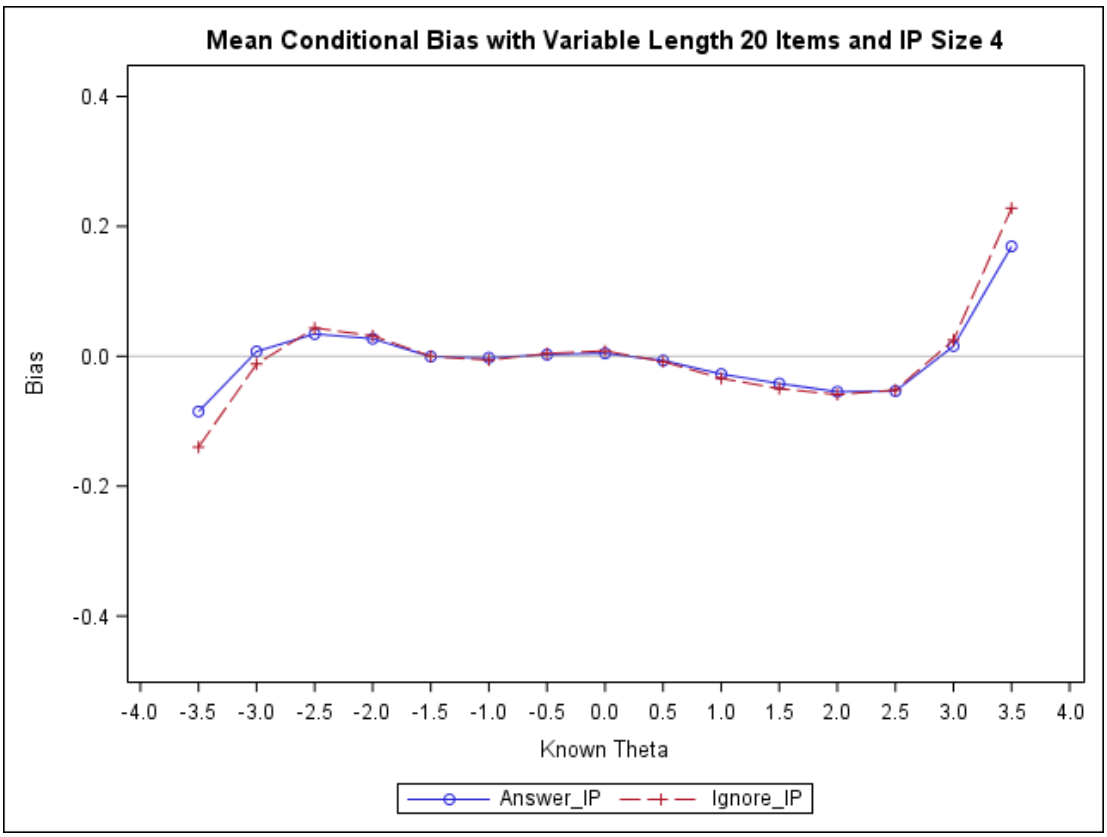

Figure 11D. Plot of Mean Bias Conditional on Known Theta for Variable Length 20 Items, IP Size 4, Forced Answer & Ignore Conditions

As expected, the conditional grand mean SEs, presented in Figure 12, 13, 14, and 15, followed the same patterns as seen with the estimated thetas and correlations. Figure 12A shows that the IP size of 4 for the fixed length 15 item FA test condition resulted in the lowest grand mean SE for all ability levels. The traditional condition resulted in the highest grand mean SE for all ability levels. However, under the Ign conditions, for all IP sizes, the resulting grand mean SE is practically the same (see Figure 12A, bottom plot). Figures 12B, 12C, and 12D display the grand mean conditional SE for the three IP sizes, which indicate the same pattern seen with the mean RMSE in the FA and Ign conditions. The FA conditions resulted in lower grand mean SEs than the Ign conditions for all ability levels, with this discrepancy increasing as IP size increased.

Increasing test length had the same result with the grand mean SE as it did with the mean RMSE. As seen in Figure 13A (top plot), the fixed length 20 item FA test condition resulted in smaller differences in grand mean SE for the IP size conditions across ability levels, with the IP size of 4 resulting in the lowest SEs. Figure 13A (bottom plot) plots the grand mean SE under the Ign test conditions, which looks similar to Figure 12A for the FA test conditions, with almost no differences in grand mean SE for all abilities. Figures 13B, 13C, and 13D show that increasing test length decreases the differences between the FA and Ign conditions in grand mean SE, for all abilities. Although the differences in mean SE decreased with increases in test length, increases in IP size resulted in decreases in grand mean SE for the FA conditions, but no effect on the Ign conditions, resulting in a widening gap between the two lines (see Figures 13B, 13C, and 13D).

The pattern seen in the fixed length conditions is seen in the variable length conditions. Figure 14A displays the grand mean SE for all IP sizes under the FA condition (top) and Ign

condition (bottom). The IP size of 4 resulted in the lowest grand mean SE for all abilities and the traditional conditions resulted in the largest mean SE. The Ign conditions resulted in conditional mean SE that is virtually identical for all IP size conditions. Again, the FA conditions resulted in lower conditional mean SEs as compared to the Ign conditions, with the SE decreasing for the FA conditions as IP size increased (see Figures 14B, 14C, and 14D). Increasing test length had the same impact on the grand mean conditional SE in the variable length conditions as it did in the fixed length conditions. As seen in Figure 15A (top plot), all IP sizes resulted in lower mean SEs across all abilities, with the IP size of 4 resulting in the lowest mean SE and the traditional resulting in the highest. The Ign conditions (see Figure 15A, bottom) shows identical mean SEs for all abilities. The same pattern is seen in Figure 15B, 15C, and 15D, with the differences in mean SE between the FA and Ign conditions shrinking as test length increased. Still, as IP size increased, the mean SE under the FA conditions decreased.

Overall, the Traditional CAT (IP size of 0) conditions resulted in the largest conditional grand mean SEs across the range of theta for all stopping rule conditions as compared to the other IP size FA conditions. Conversely, the traditional CAT (IP size of 0) conditions resulted in very similar conditional grand mean SEs compared to that of all IP size Ignore conditions across the range of known thetas. When test length increased from 15 to 20 with both fixed and variable length tests, the conditional grand mean SEs in the Ignore conditions converged, resulting in the same SEs for all known theta values. When IP size increased, the differences between conditional grand mean SEs in the Forced Answer and Ignore conditions increased, with the Forced Answer conditions resulting in lower mean SEs across the range of known thetas.

The most precise measurement was of the abilities in the center of the ability distribution. The peak of the test information function was at $\theta = -0.6$, which is the ability at which the test

measures most precisely. This is evident in the plots of mean SEs conditional on known theta (see Figure 12A, 13A, 14A, and 15A), with the lowest SEs falling between $\theta = -1.0$ to $\theta = 0$ across all conditions. As ability decreased below $\theta = -2$, mean conditional SEs also increased. The same pattern is seen on the positive extreme of the ability continuum.

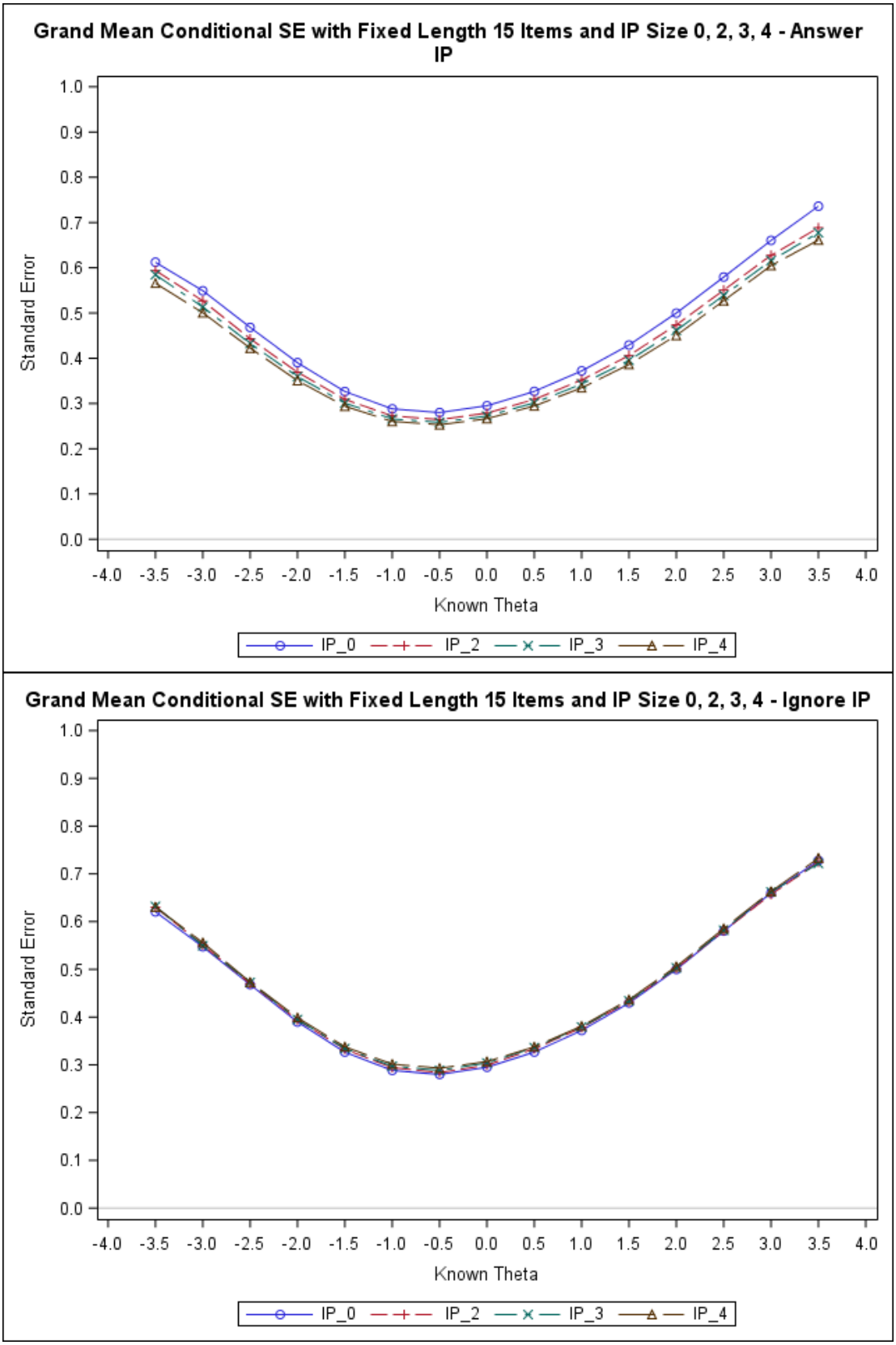

Figure 12A. Plots of Mean Standard Error (SE) Conditional on Known Theta for Fixed Length 15 Items, IP Size 0, 2, 3, & 4, Forced Answer & Ignore Conditions

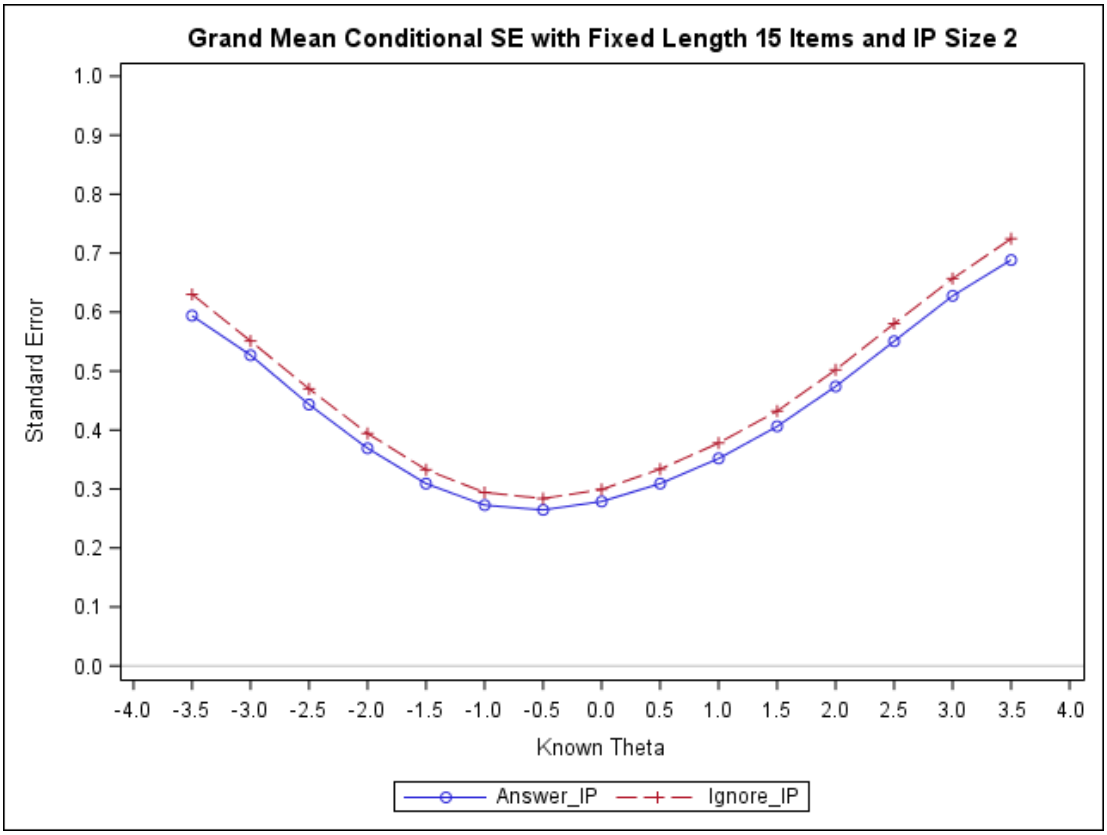

Figure 12B. Plot of Mean Standard Error (SE) Conditional on Known Theta for Fixed Length 15 Items, IP Size 2, Forced Answer & Ignore Conditions

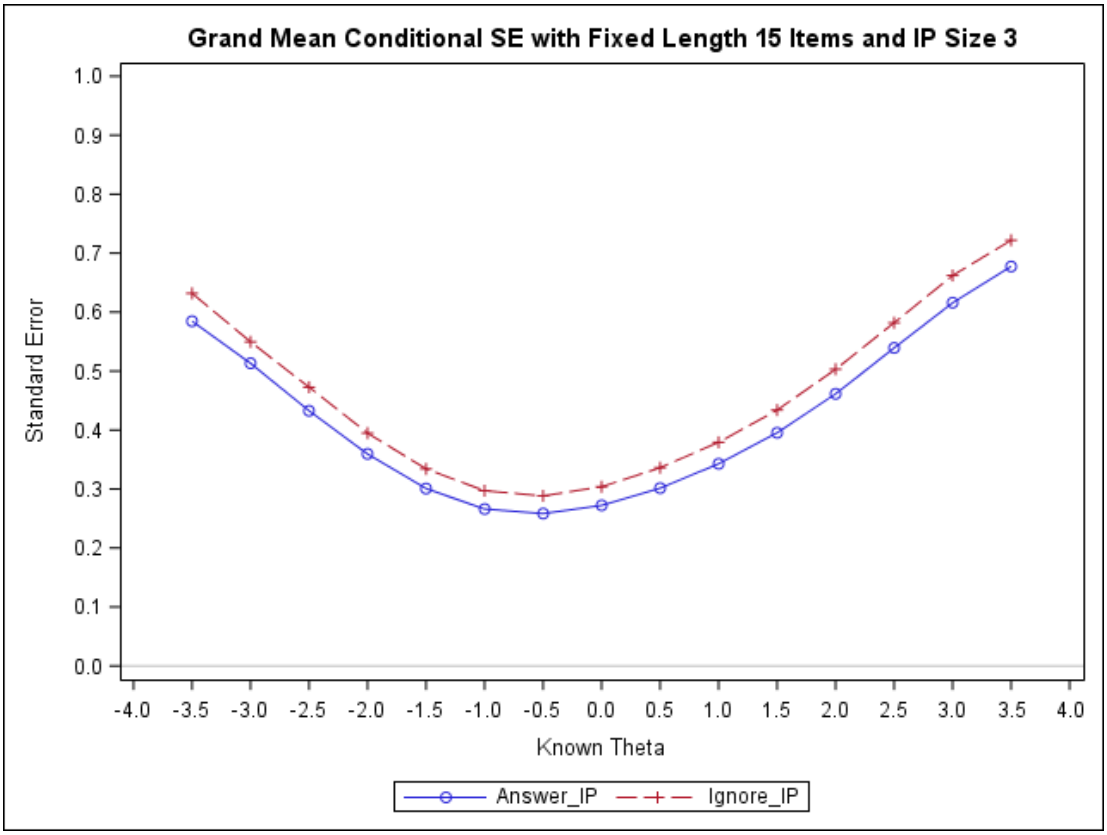

Figure 12C. Plot of Mean Standard Error (SE) Conditional on Known Theta for Fixed Length 15 Items, IP Size 3, Forced Answer & Ignore Conditions

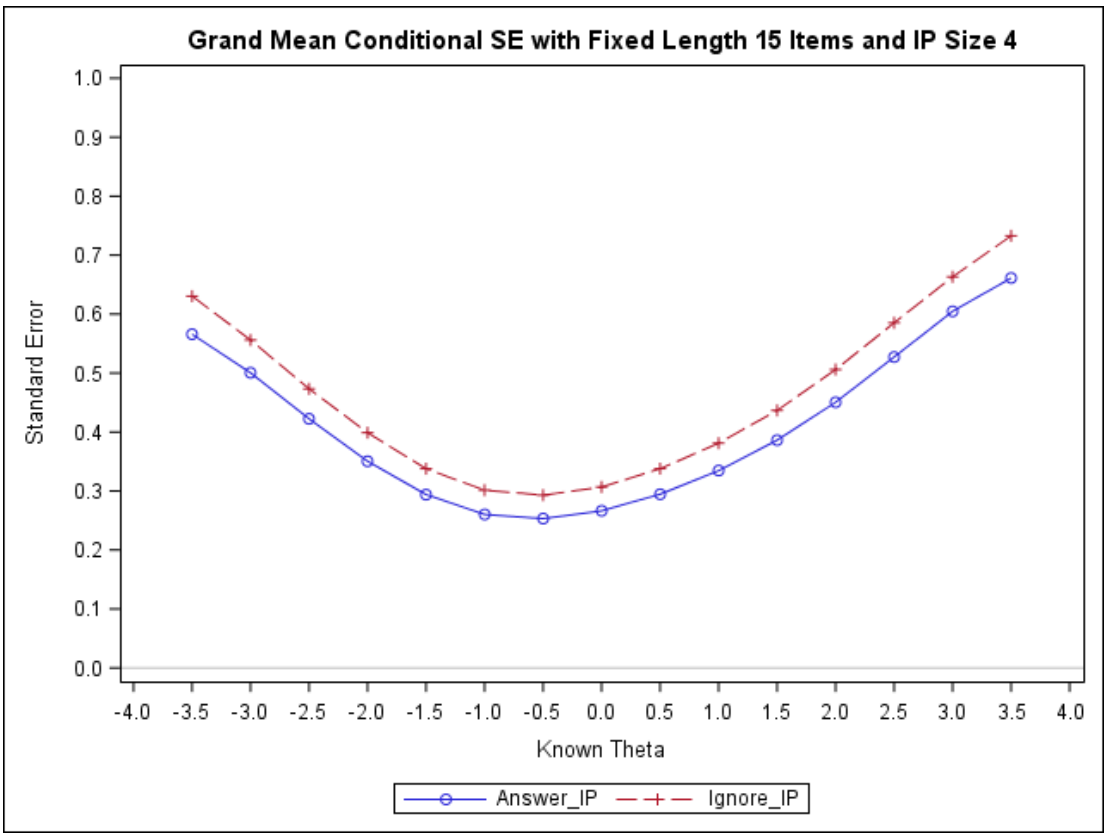

Figure 12D. Plot of Mean Standard Error (SE) Conditional on Known Theta for Fixed Length 15 Items, IP Size 4, Forced Answer & Ignore Conditions

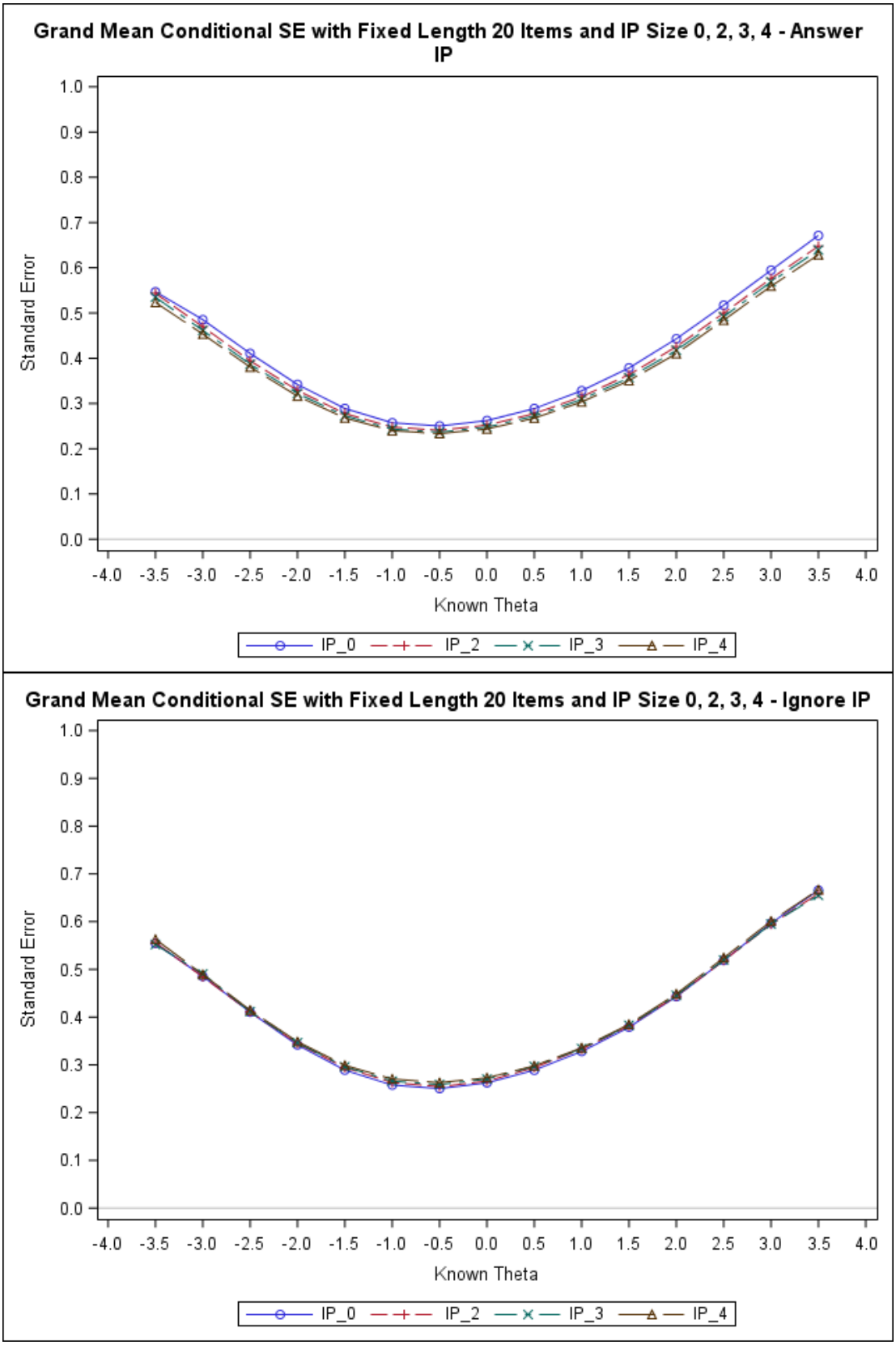

Figure 13A. Plots of Mean Standard Error (SE) Conditional on Known Theta for Fixed Length 20 Items, IP Size 0, 2, 3, & 4, Forced Answer & Ignore Conditions

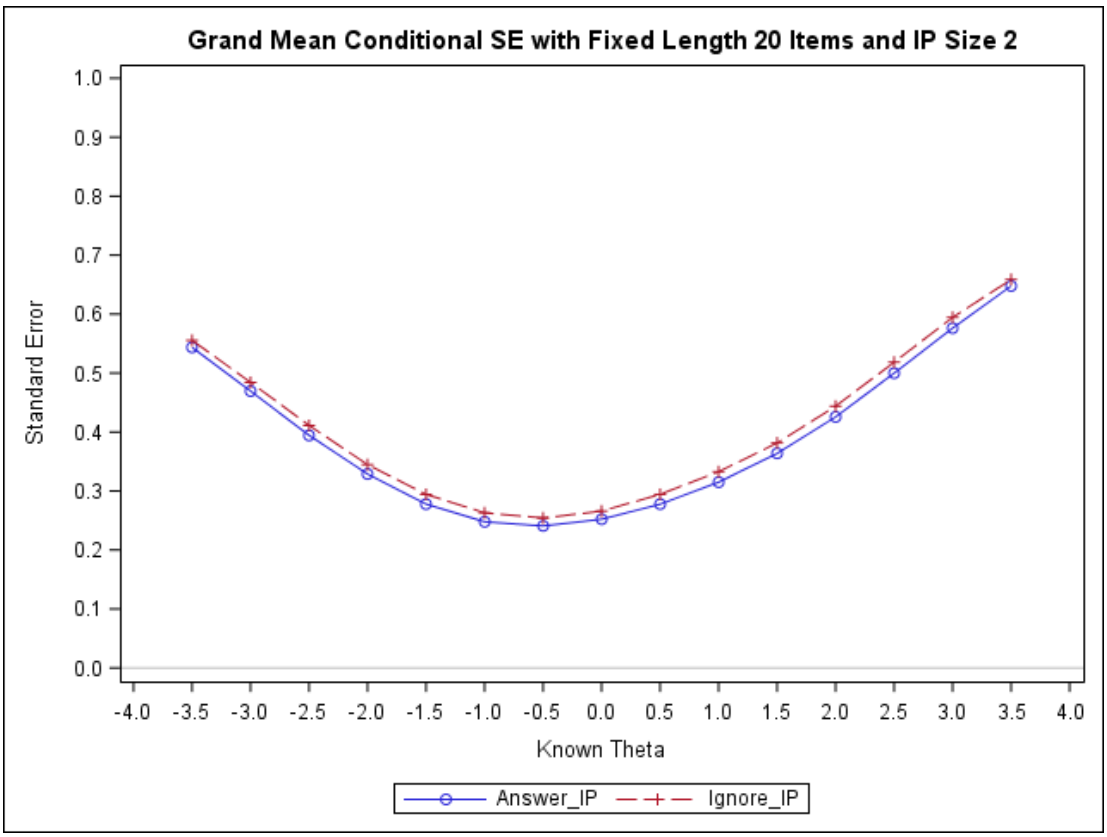

Figure 13B. Plot of Mean Standard Error (SE) Conditional on Known Theta for Fixed Length 20 Items, IP Size 2, Forced Answer & Ignore Conditions

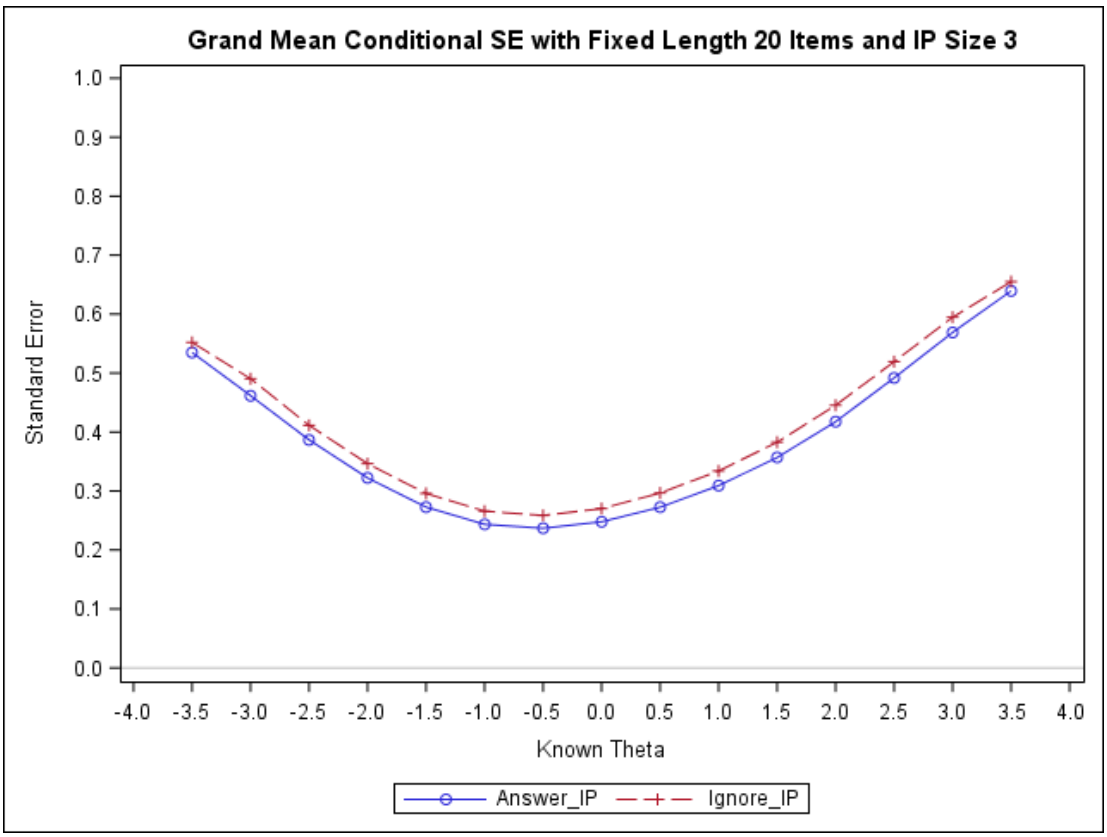

Figure 13C. Plot of Mean Standard Error (SE) Conditional on Known Theta for Fixed Length 20 Items, IP Size 3, Forced Answer & Ignore Conditions

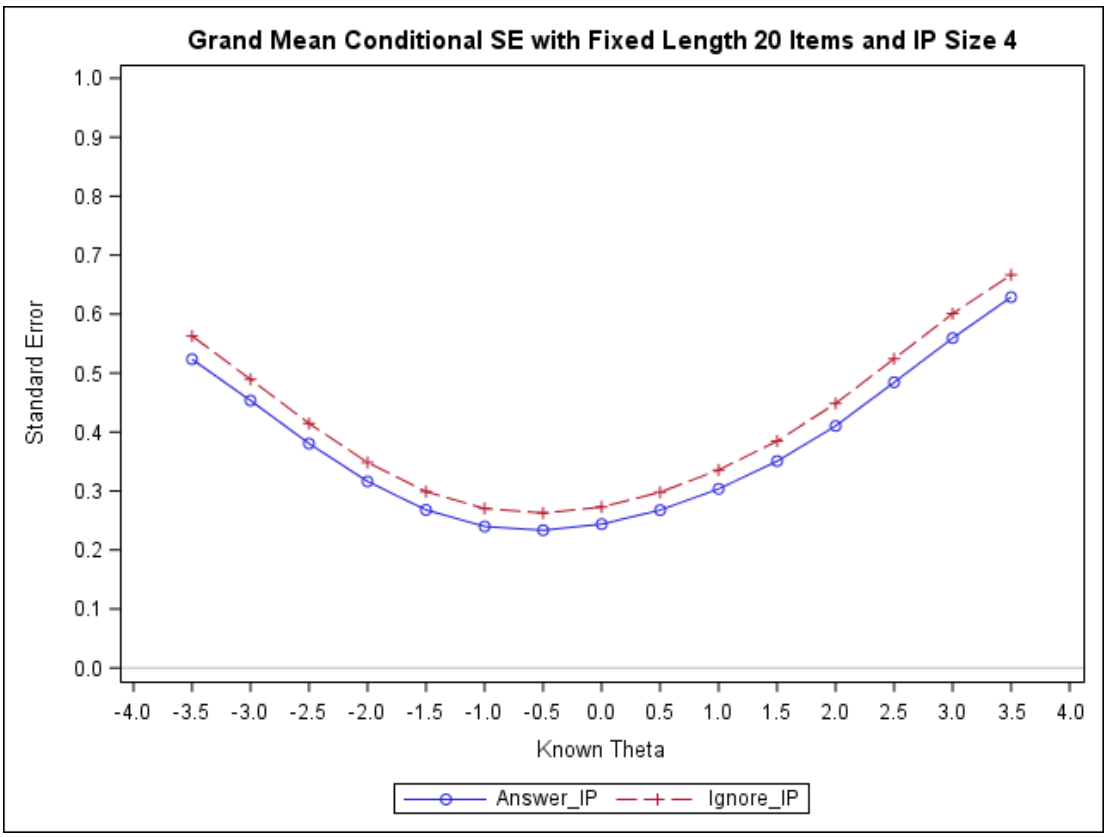

Figure 13D. Plot of Mean Standard Error (SE) Conditional on Known Theta for Fixed Length 20 Items, IP Size 4, Forced Answer & Ignore Conditions

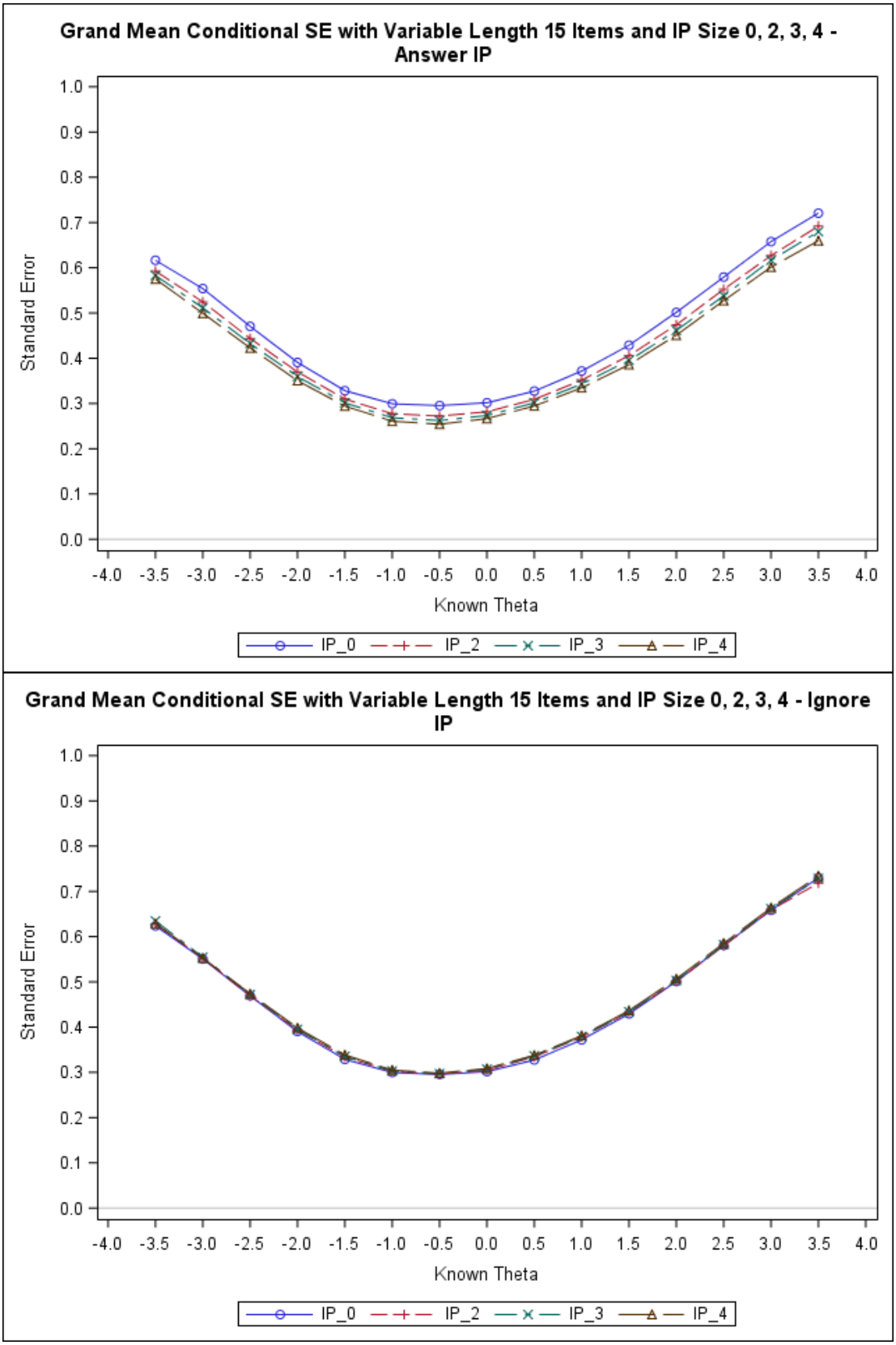

Figure 14A. Plots of Mean Standard Error (SE) Conditional on Known Theta for Variable Length 15 Items, IP Size 0, 2, 3, & 4, Forced Answer & Ignore Conditions

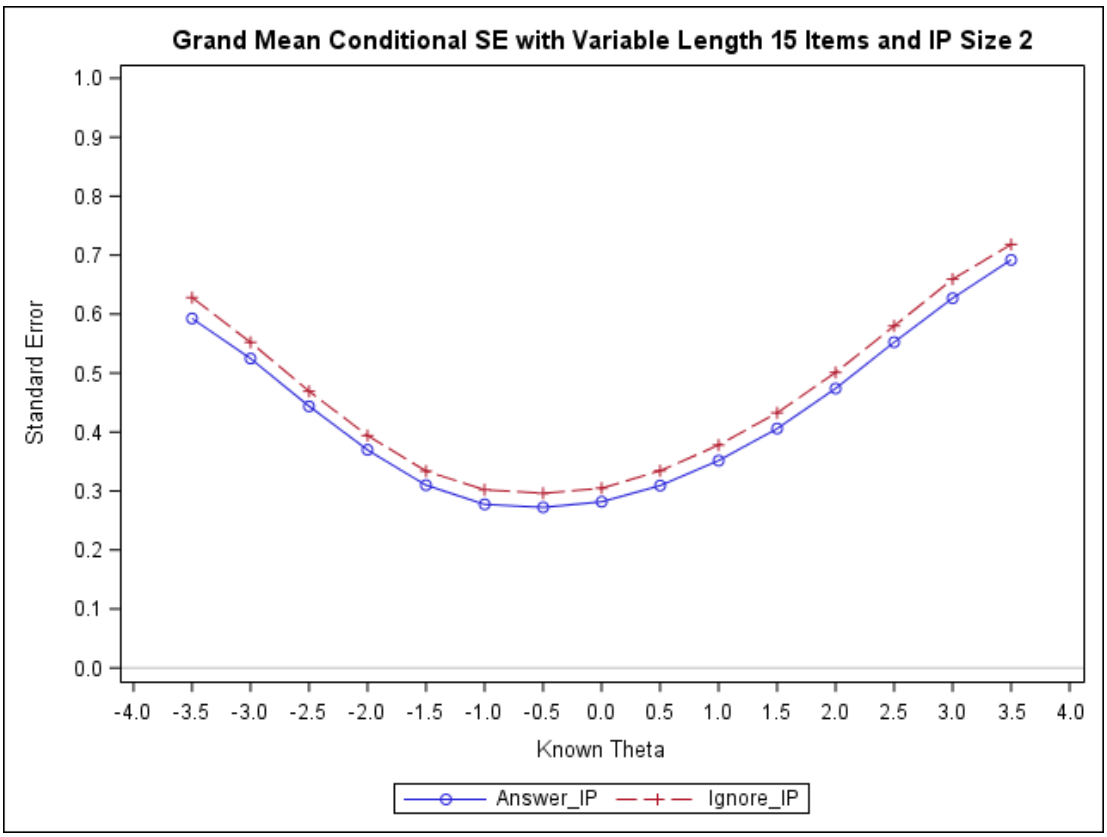

Figure 14B. Plot of Mean Standard Error (SE) Conditional on Known Theta for Variable Length 15 Items, IP Size 2, Forced Answer & Ignore Conditions

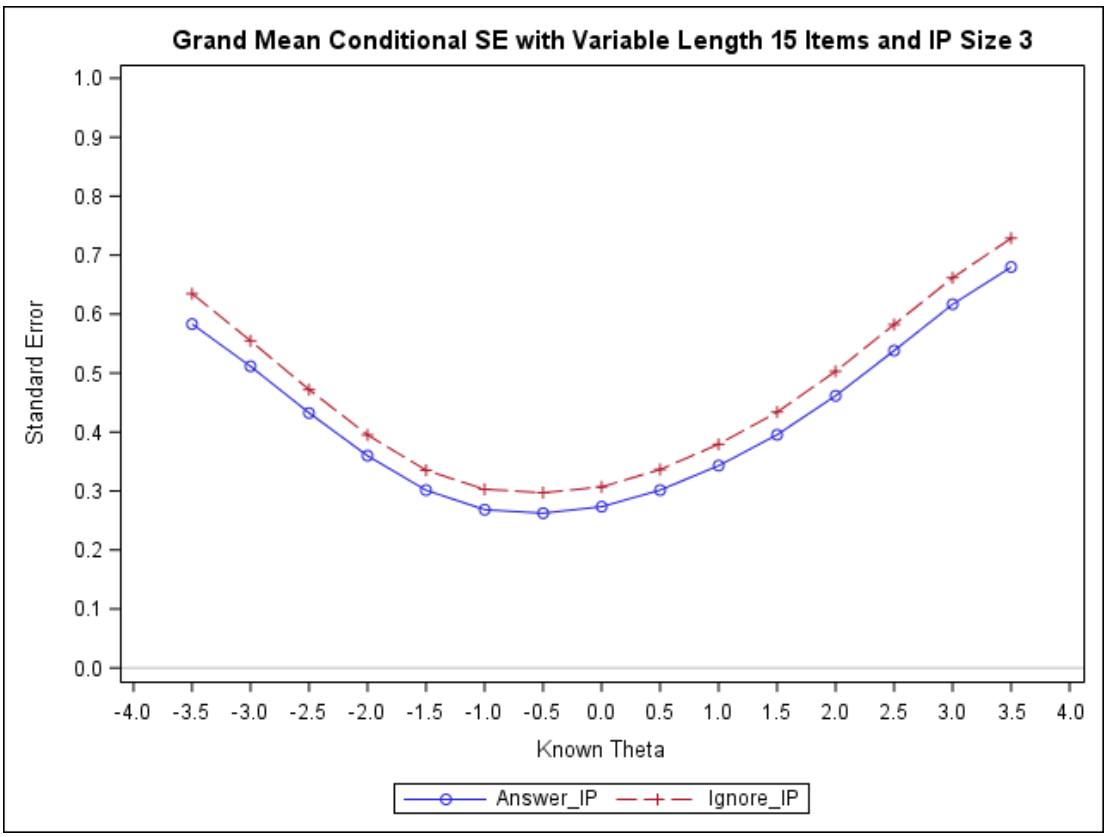

Figure 14C. Plot of Mean Standard Error (SE) Conditional on Known Theta for Variable Length 15 Items, IP Size 3, Forced Answer & Ignore Conditions

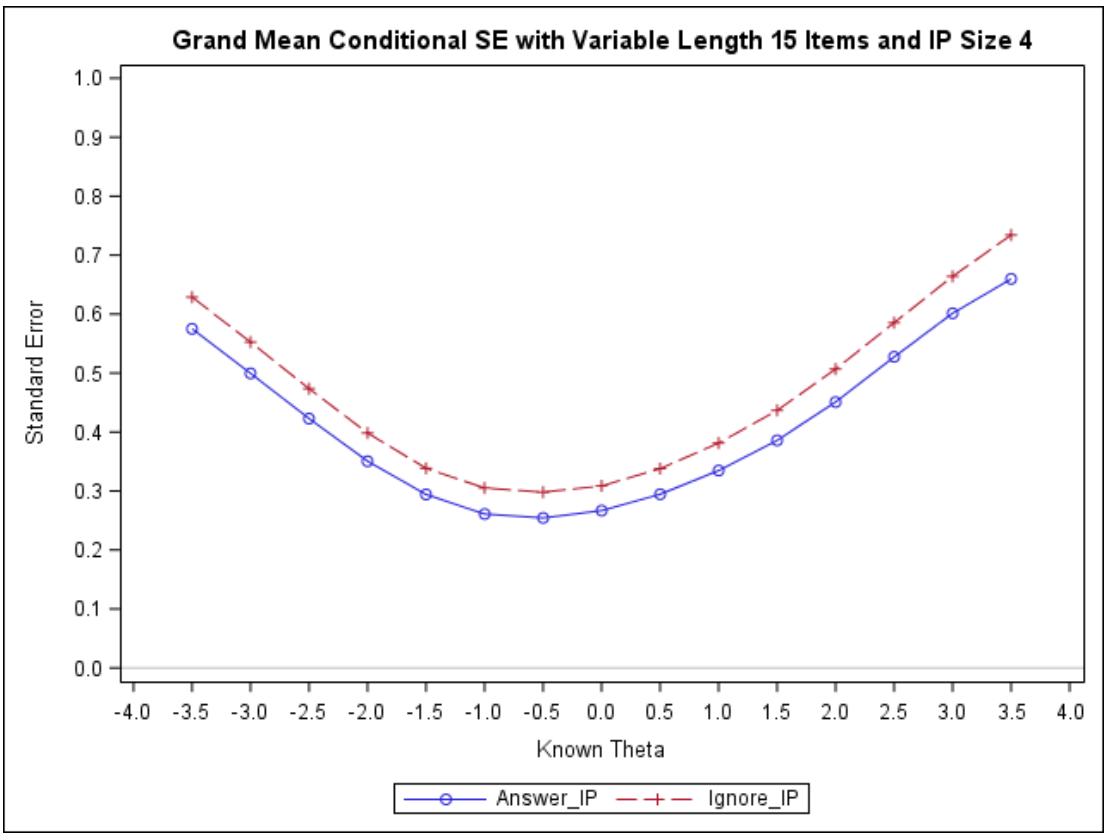

Figure 14D. Plot of Mean Standard Error (SE) Conditional on Known Theta for Variable Length 15 Items, IP Size 4, Forced Answer & Ignore Conditions

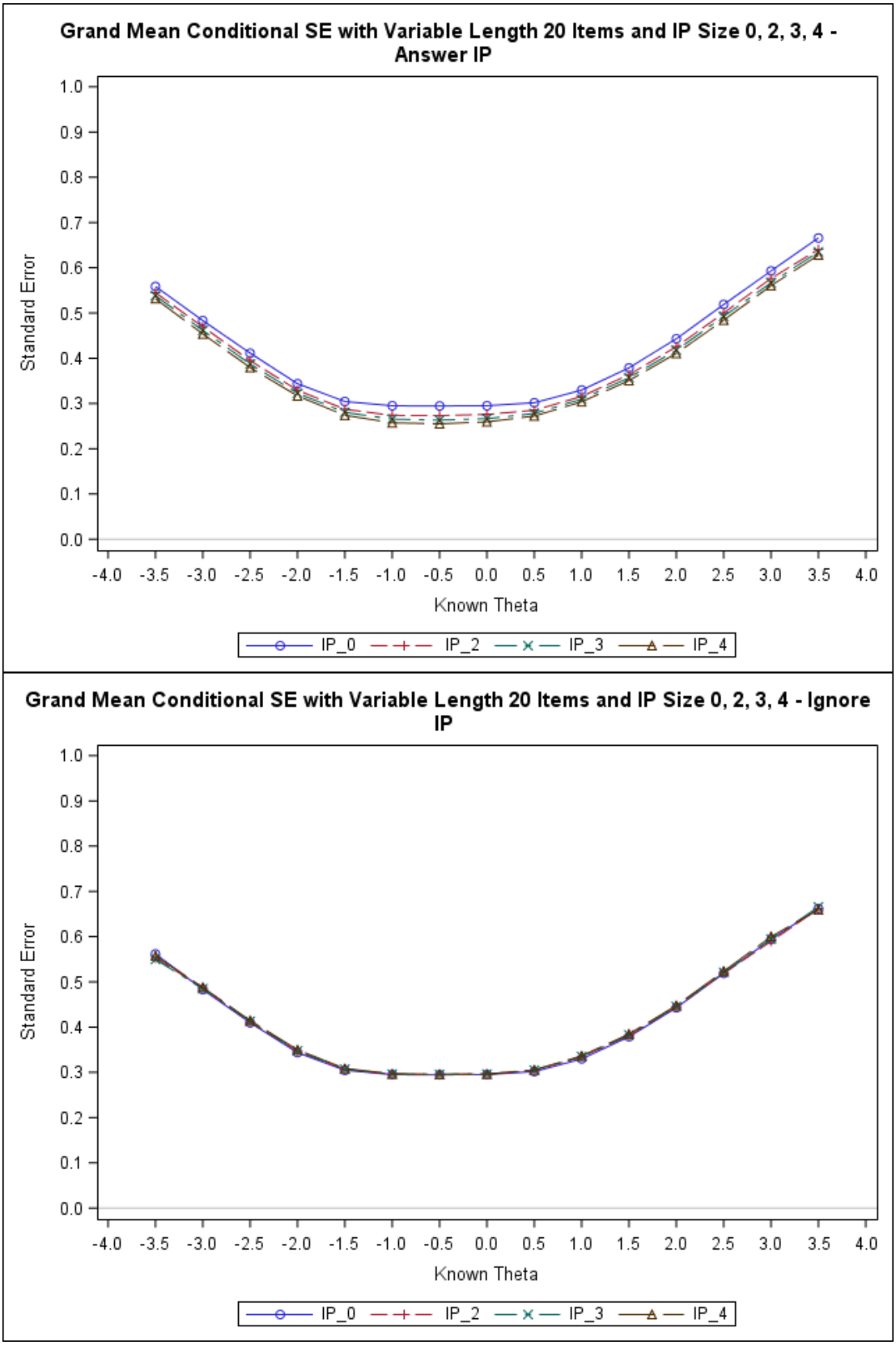

Figure 15A. Plots of Mean Standard Error (SE) Conditional on Known Theta for Variable Length 20 Items, IP Size 0, 2, 3, & 4, Forced Answer & Ignore Conditions

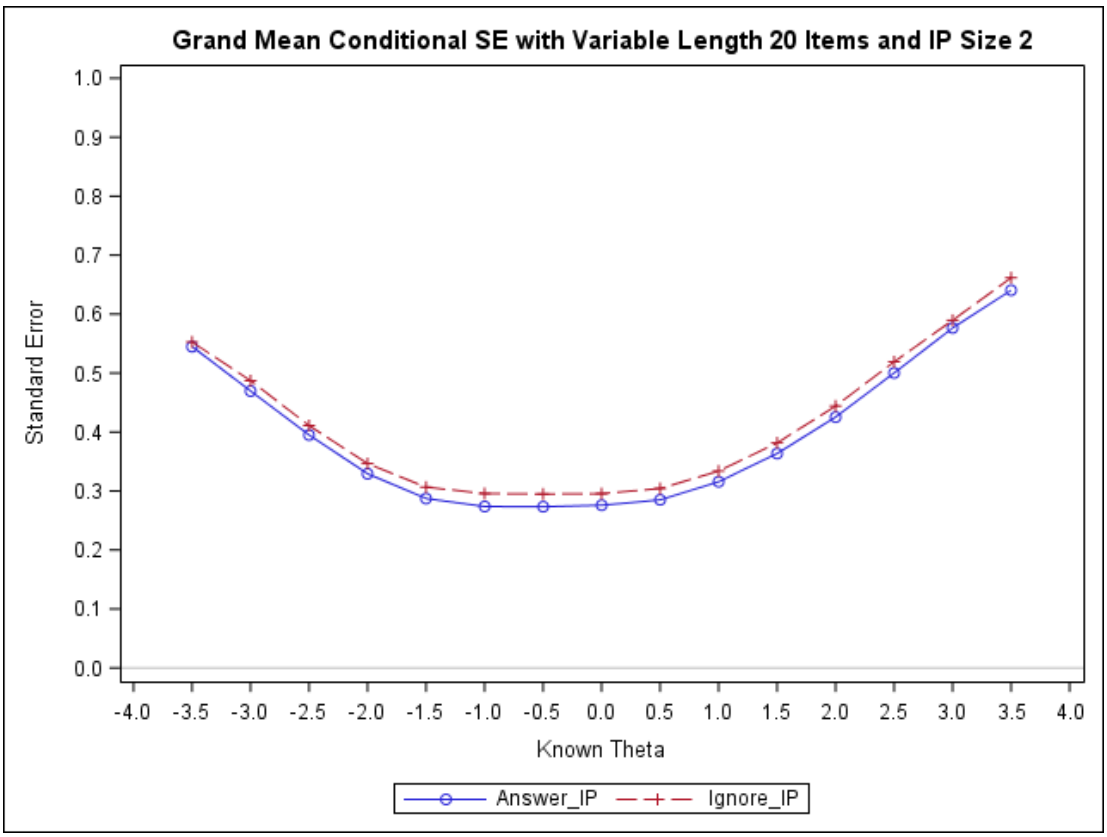

Figure 15B. Plot of Mean Standard Error (SE) Conditional on Known Theta for Variable Length 20 Items, IP Size 2, Forced Answer & Ignore Conditions

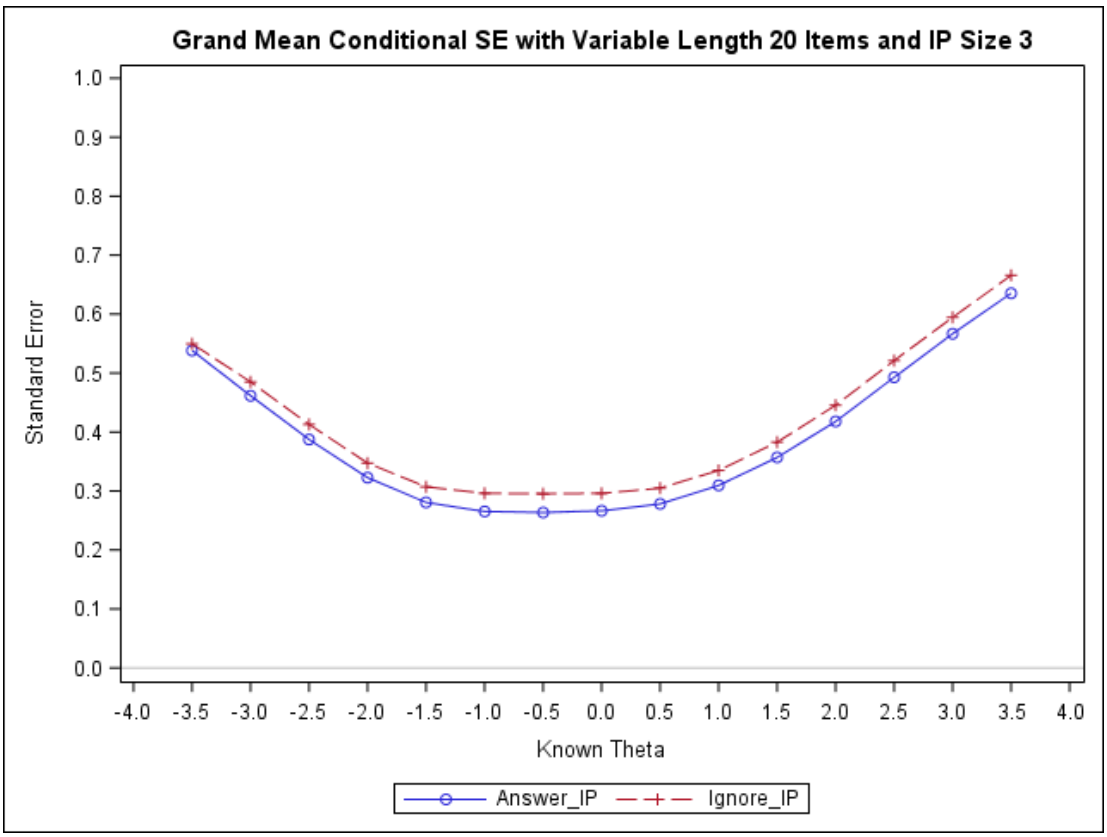

Figure 15C. Plot of Mean Standard Error (SE) Conditional on Known Theta for Variable Length 20 Items, IP Size 3, Forced Answer & Ignore Conditions

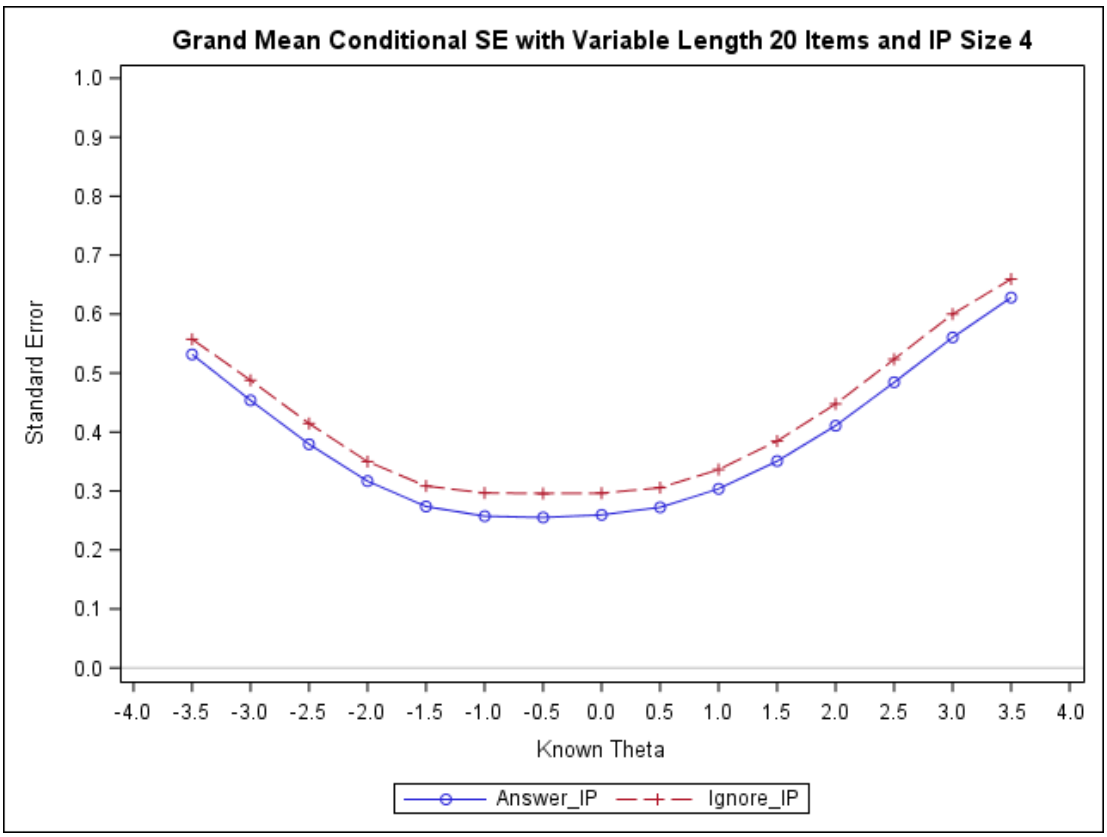

Figure 15D. Plot of Mean Standard Error (SE) Conditional on Known Theta for Variable Length 20 Items, IP Size 4, Forced Answer & Ignore Conditions

Item Pocket Usage

A portion of the evaluation of the IP method is assessment of the IP usage, although this was not explicitly asked in the research questions. The use of the IP is assumed to follow the same pattern seen in Han's (2013) study in which the use of the IP will increase as the IP size increases and ability decreases. Thus, evaluation of this pattern is of interest. This was accomplished with descriptive statistics, which are presented in Table 7. Included in Table 7 is the mean number of items placed in the pocket in each IP method condition, as well as the minimum and maximum number of items placed in the item pocket across the 500 replications. It is important to note that these numbers are higher than the IP size due to the rotation of items in and out of the item pocket throughout the course of the simulated test.

As can be seen in Table 7, as test length and IP size increased, the mean number of items placed in the pocket increased. The IP size of 2 fixed length 15 item test conditions resulted in mean IP use of 9.94 and 9.95 items for the FA and Ign conditions, respectively. Increasing the test length to 20 items resulted in mean IP use of 12.95 and 12.96 items for the FA and Ign conditions, respectively. The variable length 15 maximum item tests with an IP size of 2 resulted in the mean number of items placed in the IP of 9.56 for the FA condition and 9.55 for the Ign condition, slightly lower than the mean in the fixed length 15 item test conditions. The variable length 20 maximum item test conditions resulted in mean IP use of 10.87 items for both the FA and Ign conditions, which is, on average, almost two less items than the fixed length 20 item test conditions.

The mean number of items placed in the IP increased with IP size increases. The IP size of 3 fixed length 15 item test conditions resulted in mean IP use of 10.69 items for both FA and Ign conditions. Increasing test length increased the mean number of items placed in the IP, with

the fixed length 20 item test conditions resulting in a mean of 13.70 items placed in the IP for both FA and Ign conditions. The IP size of 3 variable length 15 maximum item tests resulted in a mean of 10.41 items placed in the IP for both FA and Ign conditions, which is slightly less than the corresponding fixed length conditions. The variable length 20 maximum item test conditions resulted in slightly higher means than the shorter tests, with 11.79 items for both FA and Ign conditions. Nonetheless, this is almost 2 items less (on average) than the corresponding fixed length test conditions.

Increasing IP size to 4 resulted in increases in the mean number of items placed in the IP. The fixed length 15 item test conditions resulted in a mean of 11.39 items placed in the IP for both FA and Ign conditions. The condition with the largest mean number of items placed in the pocket is the IP size of 4 with the fixed length 20 item test condition, with a mean of 14.41 items for both FA and Ign conditions. The variable length 15 maximum item tests again resulted in slightly lower mean number of items placed in the IP compared to the fixed length test, with a mean of 11.21 items and 11.22 items for the FA and Ign conditions, respectively. The variable length 20 maximum item tests with an IP size of 4 resulted in mean IP use of 12.67 and 12.68 for the FA and Ign conditions, respectively.

Generally, the Forced Answer and Ignore conditions resulted in the same mean number of items placed in the pocket. This was expected but was assessed for program validity purposes. The minimum number of items placed in the pocket for all conditions was 1, except for the IP size of 3 with the fixed length 20 item test condition in which items in the pocket were ignored and the IP size of 4 with the fixed length 20 item test condition in which examinees were either forced to answer or ignore items in the pocket, which resulted in a minimum of 2 items placed in the pocket. On average, the maximum number of items placed in the pocket increased

with test length and IP size increases. The IP size of 4 with the fixed length 20 item test and the IP size of 4 with the variable length 20 item test for both item completion conditions resulted in the highest maximum number of items placed in the pocket throughout the test with 23 items. The maximum number of items placed in the pocket follows the same pattern seen with the mean, which increased with test length and IP size increases.

Condition			Item Pocket Usage		
			Mean	Min	Max
IP Size 2	Fixed 15 Items	Forced Answer	9.94	1	16
		Ignored	9.95	1	16
	Fixed 20 Items	Forced Answer	12.95	1	21
		Ignored	12.96	1	21
	Variable 15 Items	Forced Answer	9.56	1	16
		Ignored	9.55	1	16
	Variable 20 Items	Forced Answer	10.87	1	21
		Ignored	10.87	1	21
IP Size 3	Fixed 15 Items	Forced Answer	10.69	1	17
		Ignored	10.69	1	17
	Fixed 20 Items	Forced Answer	13.70	1	21
		Ignored	13.70	2	22
	Variable 15 Items	Forced Answer	10.41	1	17
		Ignored	10.41	1	17
	Variable 20 Items	Forced Answer	11.79	1	22
		Ignored	11.79	1	22
IP Size 4	Fixed 15 Items	Forced Answer	11.39	1	18
		Ignored	11.39	1	18
	Fixed 20 Items	Forced Answer	14.41	2	23
		Ignored	14.41	2	23
	Variable 15 Items	Forced Answer	11.21	1	18
		Ignored	11.22	1	18
	Variable 20 Items	Forced Answer	12.67	1	23
		Ignored	12.68	1	23

Table 7. Mean, Minimum, and Maximum Number of Items Placed in the Item Pocket Averaged Across Replications

Conditional Item Pocket Usage

The conditional item pocket use is of interest due to the likely use of the pocket varying on ability level, as was seen in Han (2013). Conditional IP usage was assessed by plotting the grand mean conditional on known theta across the range of θ from -3.5 to +3.5 with 0.5 increments, averaged across the 500 replications. These plots are presented in Figures 16 through 19. Figure 16A displays the mean item pocket use for the fixed length 15 item test under the FA condition, which shows that the average number of items placed in the pocket increased for IP sizes of 2, 3, and 4 for abilities at and below θ = 0. In addition, for the full range of abilities, IP use generally increased with IP size increases. This same general pattern is seen under the Ign condition for the fixed length 15 item tests with IP sizes of 2, 3, and 4 (see Figure 16B). Increasing test length to 20 items for the fixed length test conditions under the FA and Ign methods resulted in the same pattern seen for the fixed length 15 item test conditions, with IP use increasing for abilities at and below θ = 0, and generally increasing with IP size increases (see Figure 17A and 17B). To summarize, generally, as ability decreased below θ = 0, IP use increased in the fixed length conditions, regardless of item completion condition. In addition, as IP size increased, IP use also generally increased.

The pattern seen with the fixed length 15 item tests is generally seen under the variable length 15 maximum item test conditions. Again, as test length increased, the IP usage increased; however, the average number of items is slightly less under the variable length 15 maximum item test conditions for abilities at and above θ = 0 for both FA and Ign conditions (see Figure 18A and 18B). Under the variable length 20 maximum item test conditions, the pattern seen thus far slightly changes. In Figure 19A, the point on the ability continuum where IP use increases, as seen in the fixed length test conditions, has shifted down to θ = -1.5. In addition, abilities above

$\theta = 0$ resulted in mean IP use slightly lower than the average seen in the corresponding fixed length test conditions. The pattern of IP use seen in the variable length 20 maximum item FA test conditions is repeated under the Ign conditions (see Figure 19B). The average number of items placed in the pocket for all ability levels generally increases as IP size increases. Interestingly, the ability level at which the average number of items placed in the pocket increases is shifted down to $\theta = -1.5$ for the variable length 20 maximum item tests under both the FA and the Ign conditions. The item completion conditions for all termination criteria resulted in similar IP use, meaning that termination criteria appeared to have no impact on IP use.

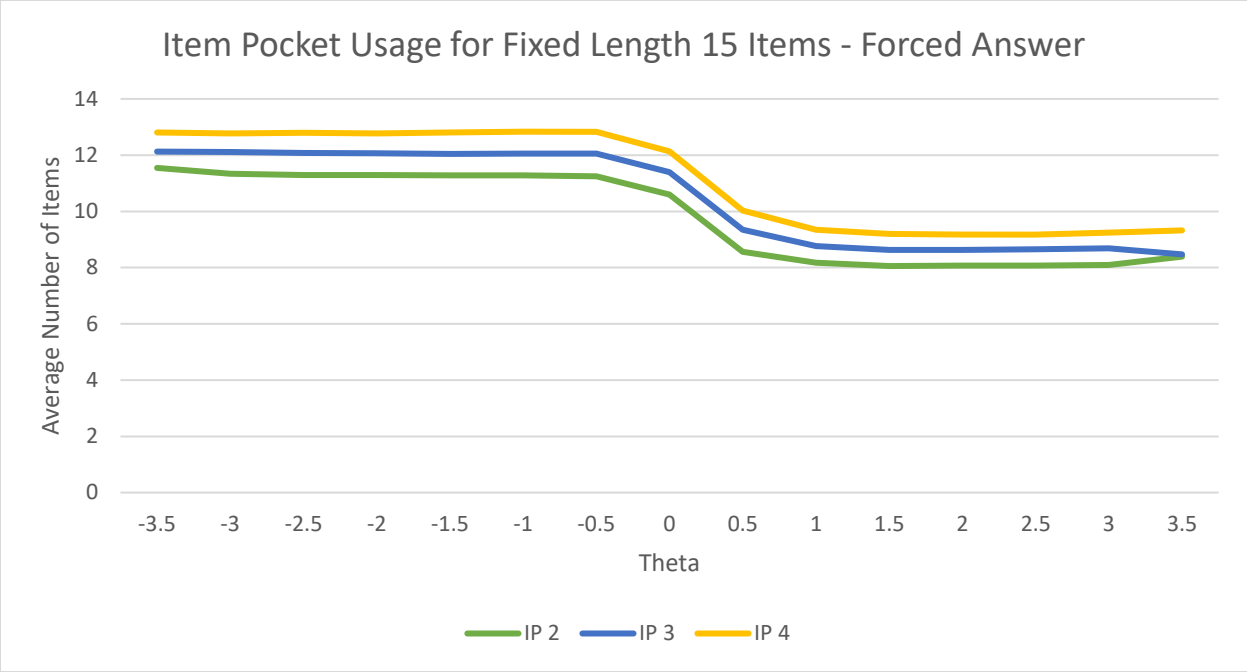

Figure 16A. Grand Mean Item Pocket Use Conditional on Known Theta, Fixed Length 15 Items, IP Size 2, 3, & 4, Forced Answer Conditions

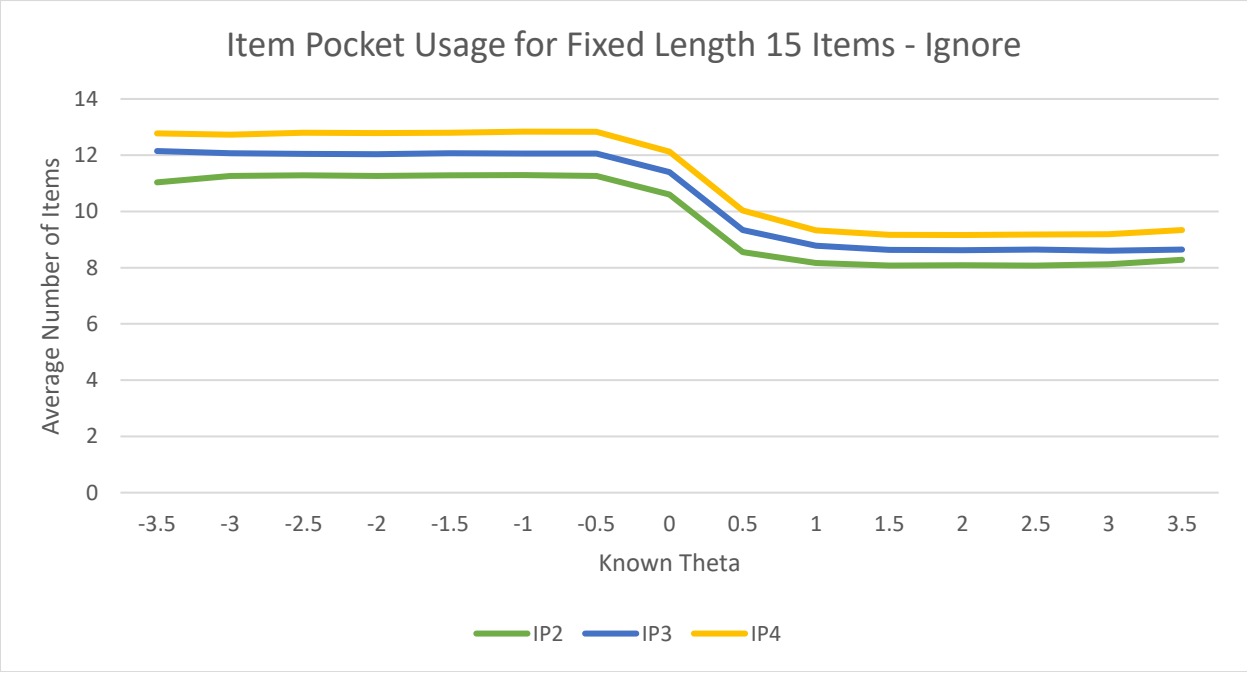

Figure 16B. Grand Mean Item Pocket Use Conditional on Known Theta, Fixed Length 15 Items, IP Size 2, 3, & 4, Ignore Conditions

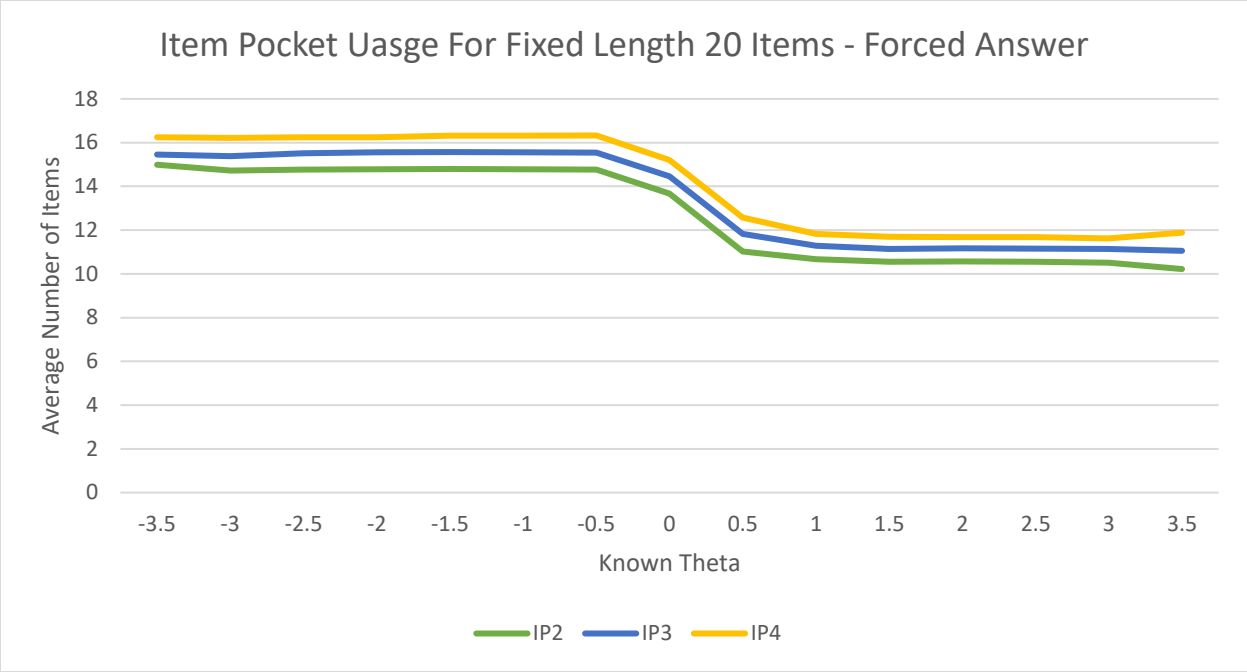

Figure 17A. Grand Mean Item Pocket Use Conditional on Known Theta, Fixed Length 20 Items, IP Size 2, 3, & 4, Forced Answer Conditions

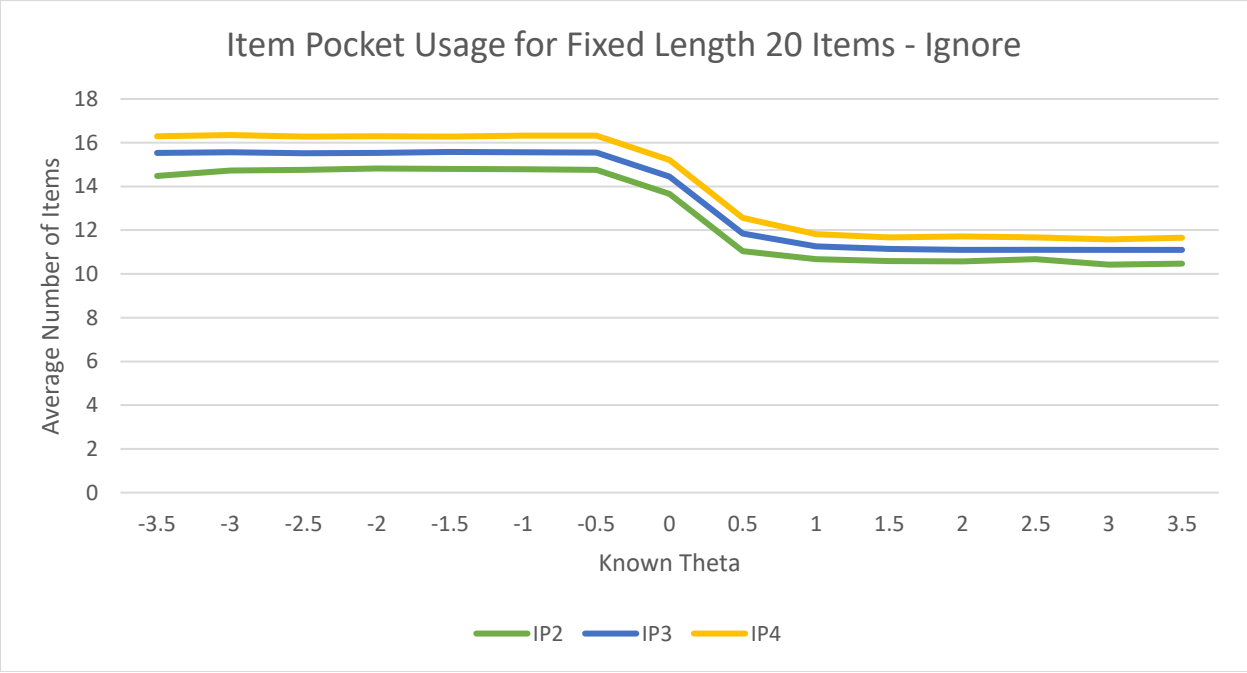

Figure 17B. Grand Mean Item Pocket Use Conditional on Known Theta, Fixed Length 20 Items, IP Size 2, 3, & 4, Ignore Conditions

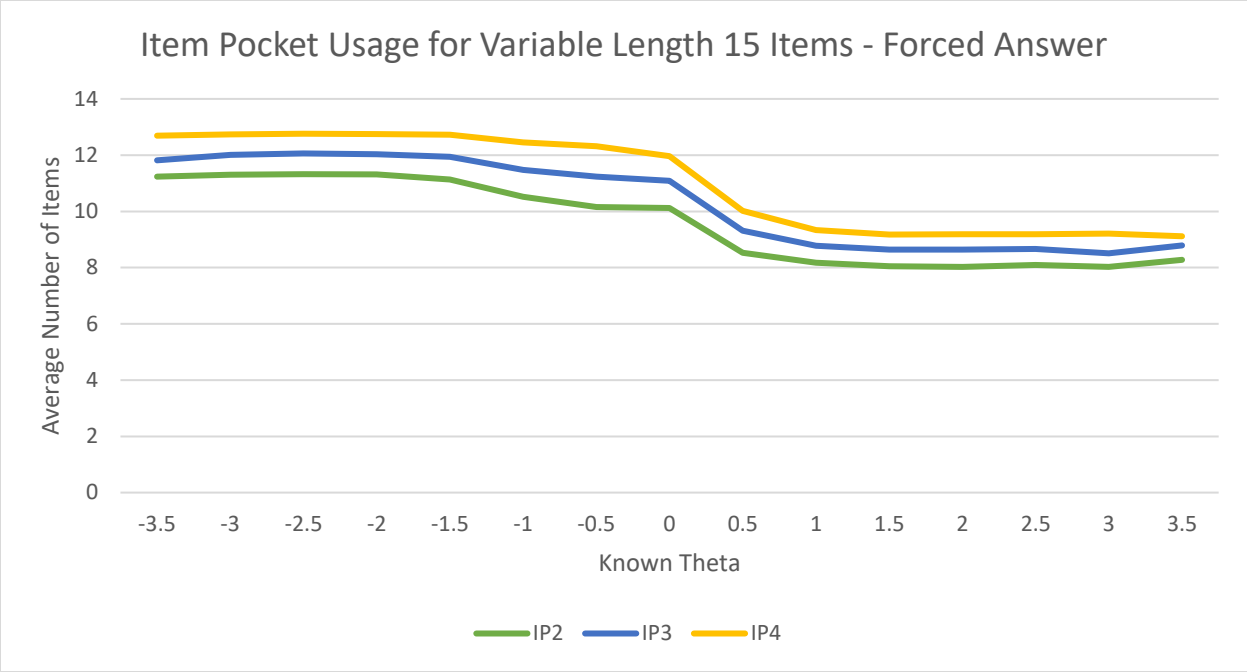

Figure 18A. Grand Mean Item Pocket Use Conditional on Known Theta, Variable Length 15 Items, IP Size 2, 3, & 4, Forced Answer Conditions

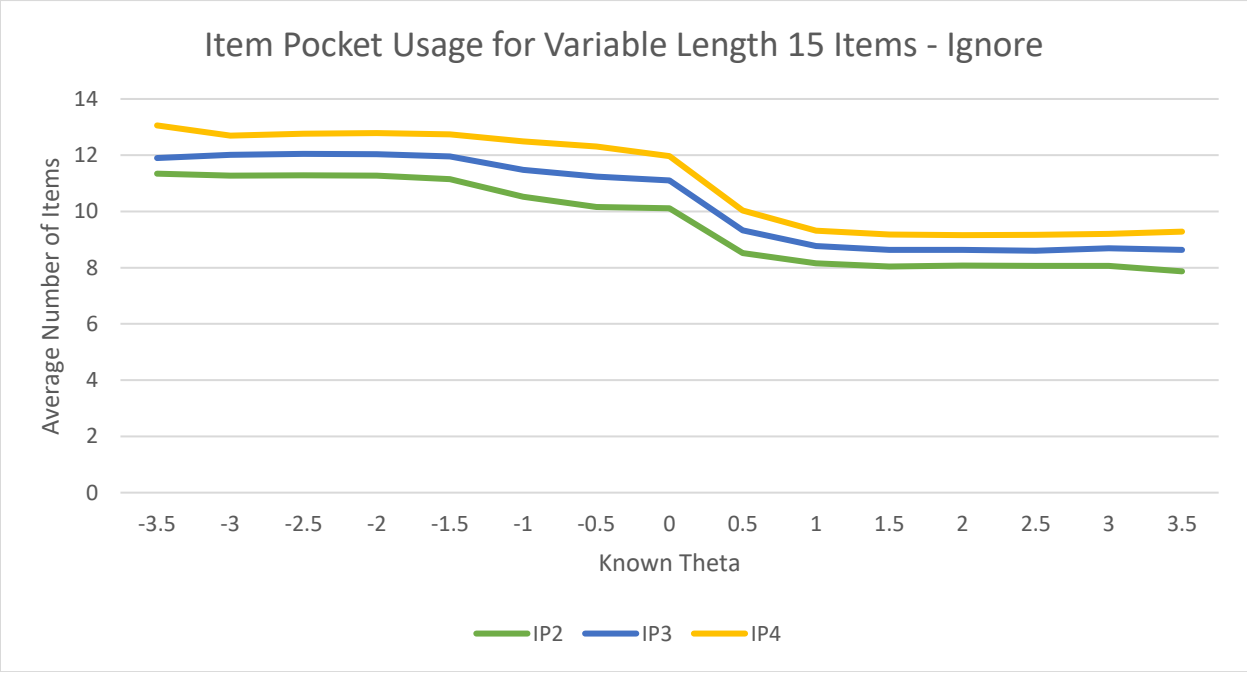

Figure 18B. Grand Mean Item Pocket Use Conditional on Known Theta, Variable Length 15 Items, IP Size 2, 3, & 4, Ignore Conditions

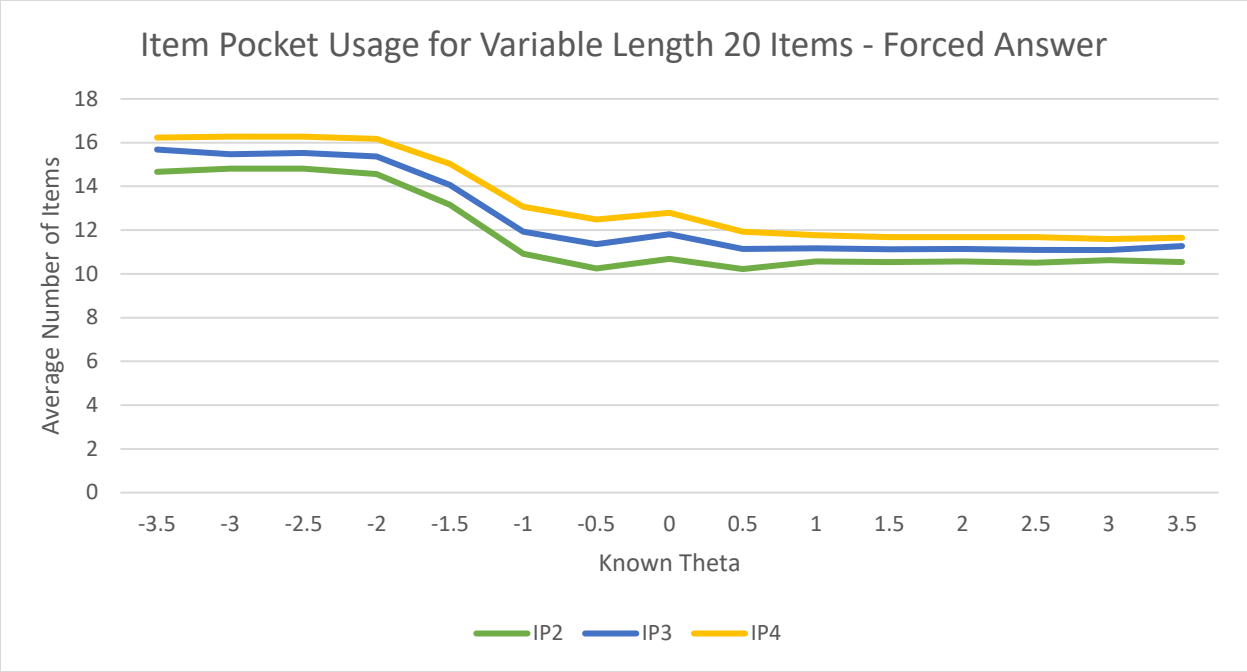

Figure 19A. Grand Mean Item Pocket Use Conditional on Known Theta, Variable Length 20 Items, IP Size 2, 3, & 4, Forced Answer Conditions

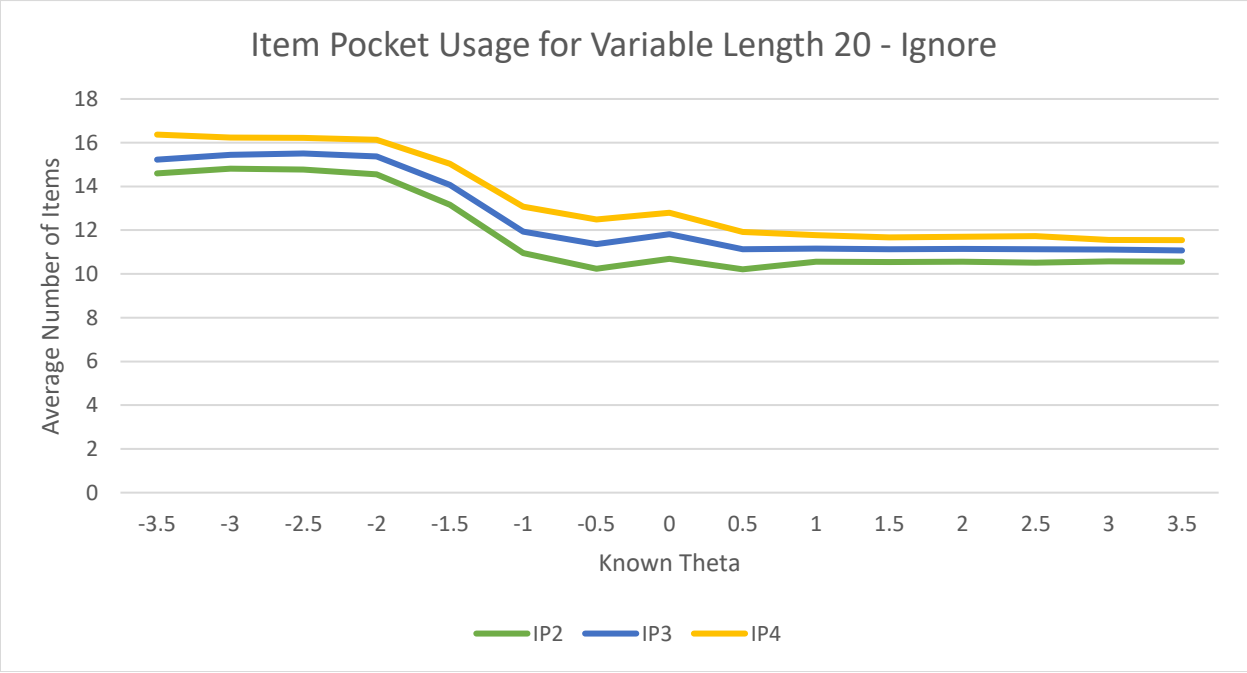

Figure 19B. Grand Mean Item Pocket Use Conditional on Known Theta, Variable Length 20 Items, IP Size 2, 3, & 4, Ignore Conditions

Test Efficiency

The efficiency of the CAT was evaluated with descriptive statistics, the mean, minimum, and maximum number of items administered (NIA), for each condition averaged across the 500 replications. The conditions with smaller mean values indicate more efficient tests. Descriptive statistics for NIA are included in Table 8. Typically, assessment of test efficiency is only evaluated when a variable length termination criteria is used. However, the Forced Answer item completion conditions included in the present study will impact the NIA and, therefore, the fixed length conditions are included. The possibility of an interaction of the IP size with the test length in variable length termination conditions was also assessed.

The mean NIA for the Traditional CAT (IP size of 0) conditions is not applicable for the fixed length tests since they all resulted in the same number of items administered. The traditional CAT (IP size of 0) variable length conditions resulted in a mean NIA of 14.27 and 16.46 for 15 item and 20 maximum items tests, respectively. In all IP size conditions, the fixed length FA tests resulted in mean NIAs slightly lower than the maximum number of items plus the IP size. For instance, the IP size of 2 FA conditions with fixed length 15 item and 20 item tests resulted in mean NIAs of 16.99 and 21.99, respectively. The maximum NIA for these conditions is 17 and 22 (i.e., 15 + 2 & 20 + 2), respectively. The mean NIA for these conditions being slightly less than the maximum is the result of some examinees having less than the maximum number of items in the pocket at the end of the test. This pattern is consistent for all of the fixed length, forced answer conditions.

The variable length test conditions with the IP method resulted in slightly higher mean NIAs as compared to the Traditional CAT (IP size of 0). This pattern holds for both item completion conditions. On average, the IP size of 2 variable length 15 item test condition

resulted in a mean of 16.55 when examinees were forced to answer items in the pocket and a mean of 14.43 when examinees ignored items in the pocket. These means are higher than the mean NIA of 14.27 found in the corresponding traditional CAT condition (IP size of 0 with a variable length 15 item test). This is explained by the administration of additional items in both item completion conditions. The Forced Answer condition requires the administration of the items in the pocket at the end of the test, which will result in a higher average NIA. The explanation of the higher mean in the Ignore condition is the loss of information from the items placed in the pocket and ultimately ignored. Skipping these items resulted in slightly more items administered on average across the 500 replications. However, this difference is a mere 0.16 average items across the 500 replications. The same pattern is seen in the IP size of 2 with a variable length 20 item test conditions, wherein the Forced Answer condition resulted in a mean NIA of 18.90 and the Ignore condition resulted in a mean NIA of 16.85, which are both higher than the mean NIA (16.46) seen in the traditional CAT (IP size of 0) with the variable length 20 item test.

As the size of the IP increased, the tests slightly lost efficiency. The IP size of 3 with variable length 15 item test conditions resulted in mean NIAs of 17.73 and 14.59 when examinees were forced to answer items in the pocket and when items in the pocket were ignored, respectively. The IP size of 3 with variable length 20 item test conditions resulted in mean NIAs of 20.17 and 17.12 when examinees were forced to answer items in the pocket and when items in the pocket were ignored, respectively. This pattern continued in the IP size of 4 conditions, with the Forced Answer conditions resulting in a higher mean NIA proportionate to the IP size, and the Ignore conditions resulting in mean NIAs only slightly higher than that of the Traditional CATs. Specifically, the IP size of 4 with fixed length 15 item tests under the forced answer

conditions resulted in a mean NIA of 18.99. Increasing test length to 20 items resulted in a mean NIA of 23.99 for the IP size of 4 in the fixed length 20 item FA test condition. The variable length 15 maximum item FA test condition resulted in a mean NIA of 18.91, whereas the Ign condition resulted in a mean NIA of 14.75, which is slightly less than the maximum number of items, but still slightly more than the IP size of 0 condition. The variable length 20 maximum item test conditions resulted in mean NIAs of 21.43 and 17.38 for the FA and Ign conditions, respectively. Again, the FA conditions resulted in mean NIA slightly higher than the maximum number of items due to the forced administration of the items remaining in the pocket and the Ign conditions resulted in mean NIAs lower than the maximum number of items, but slightly higher than the corresponding traditional IP size of 0 conditions.

Condition	Number of Items Administered

			Mean	Min	Max
Traditional (IP=0)	Fixed 15 Items		15.00	15	15
	Fixed 20 Items		20.00	20	20
	Variable 15 Items		14.27	11	15
	Variable 20 Items		16.46	11	20
IP Size 2	Fixed 15 Items	Forced Answer	16.99	15	17
		Ignored	15.00	15	15
	Fixed 20 Items	Forced Answer	21.99	20	22
		Ignored	20.00	20	20
	Variable 15 Items	Forced Answer	16.55	14	17
		Ignored	14.43	12	15
	Variable 20 Items	Forced Answer	18.90	12	22
		Ignored	16.85	12	20
IP Size 3	Fixed 15 Items	Forced Answer	17.99	15	18
		Ignored	15.00	15	15
	Fixed 20 Items	Forced Answer	22.99	21	23
		Ignored	20.00	20	20
	Variable 15 Items	Forced Answer	17.73	14	18
		Ignored	14.59	12	15
	Variable 20 Items	Forced Answer	20.17	15	23
		Ignored	17.12	12	20
IP Size 4	Fixed 15 Items	Forced Answer	18.99	15	19
		Ignored	15.00	15	15
	Fixed 20 Items	Forced Answer	23.99	22	24
		Ignored	20.00	20	20
	Variable 15 Items	Forced Answer	18.91	15	19
		Ignored	14.75	12	15
	Variable 20 Items	Forced Answer	21.43	13	24
		Ignored	17.38	12	20

Table 8. Mean, Minimum, and Maximum Number of Items Administered (NIA) Averaged Across Replications

The possibility of an interaction occurring with the implementation of the IP method in conjunction with the variable termination criteria was assessed. However, due to the large sample size, any test of this interaction will produce statistically significant results for very small differences. Therefore, the mean NIA across the 500 replications was plotted for all IP sizes, test lengths, and item completion methods in the variable length conditions in Figure 20 (A & B). As can be seen in Figures 20A and 20B, for both the FA and Ign item completion methods, the differences between the mean NIAs for the 15 item and 20 item tests at each IP size is comparable, indicative of no interaction. As IP size increased, the mean NIA increased proportionally.

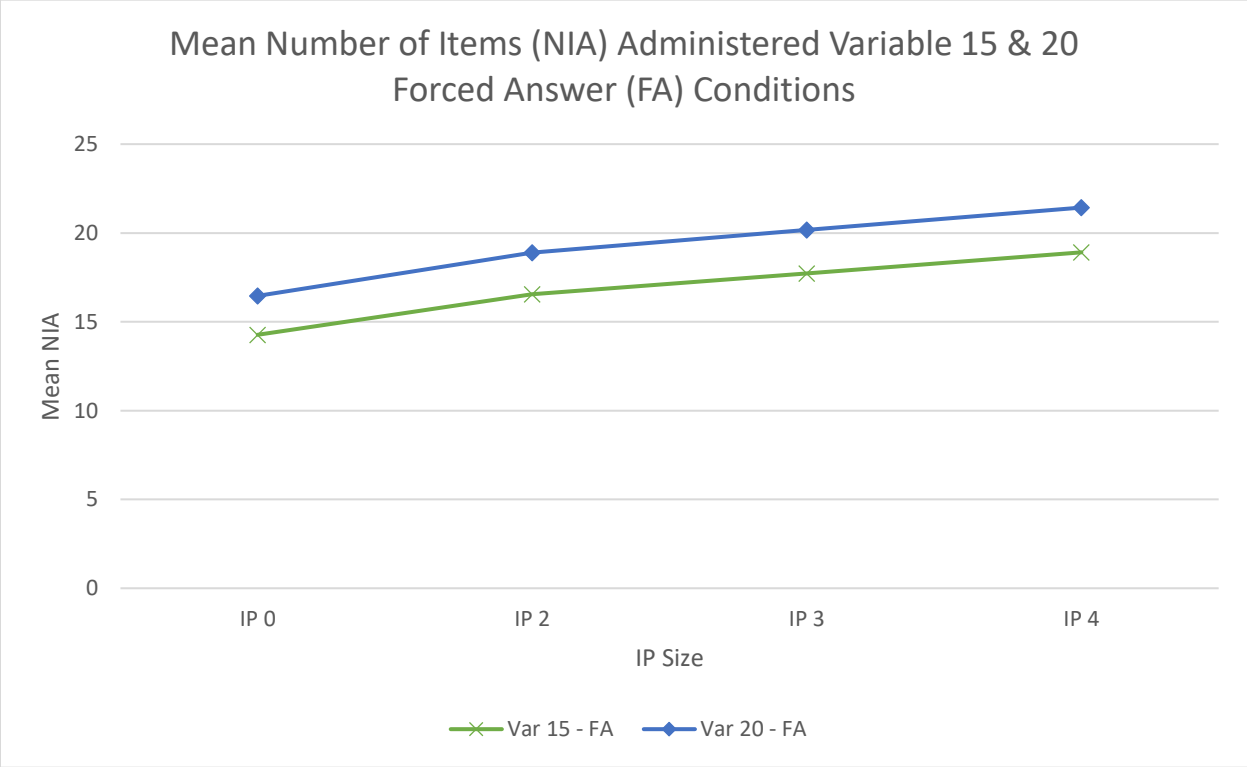

Figure 20A. Plot of Mean Number of Items Administered (NIA) for Variable Length 15 & 20 Items, IP Size 0, 2, 3, & 4, Forced Answer Conditions

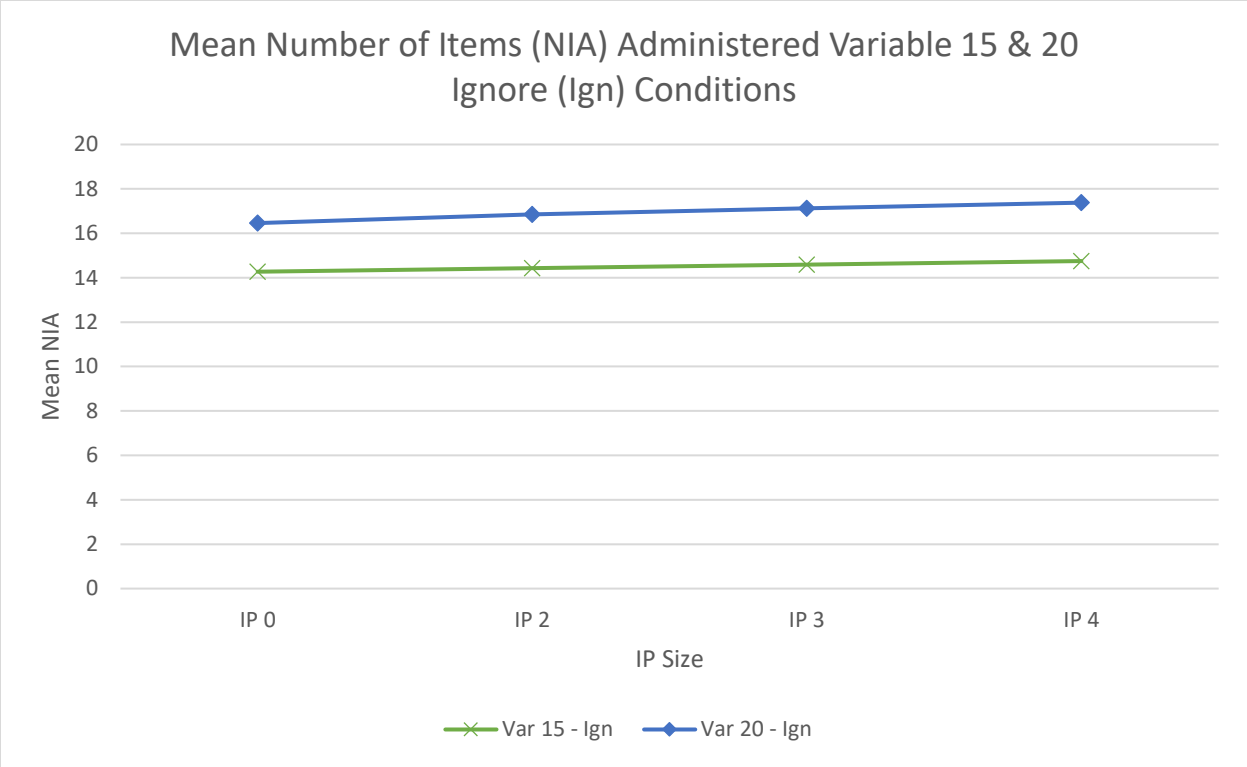

Figure 20B. Plot of Mean Number of Items Administered (NIA) for Variable Length 15 & 20 Items, IP Size 0, 2, 3, & 4, Ignore Conditions

Conditional Test Efficiency

Test efficiency is influenced by the composition of the item pool and the examinee's ability, with more efficient tests for examinees whose ability is matched to the item pool distribution. In these situations, there are many items that accurately measure those ability levels, therefore resulting in shorter tests. Examinees with abilities in the extremes of the item pool distribution will be less efficiently measured, meaning that more items are administered to those examinees because the SE criteria is never met. Evaluation of the conditional efficiency allows for the assessment of test efficiency across the range of known thetas.

The grand mean NIA averaged across the 500 replications, conditional on known theta, with 0.5 increments across the range of thetas for all of the variable length conditions are

presented in Figures 21 and 22. As expected, the variable length 15 item tests with both Forced Answer and Ignore item completion methods resulted in more efficient tests for those examinees whose abilities were between Theta of -1.5 and 0.0, with grand mean NIAs less than the maximum number of 15 items, particularly for IP sizes of 0, 2, and 3 for these abilities (see Figures 21A and 21B). However, as IP size increased, for all other abilities, this efficiency is lost, particularly for the Forced Answer conditions. This is explained by the requirement of answering the items in the pocket at the end of the test, thereby increasing the NIA. Figure 21C displays the conditional grand mean NIA for the variable length 15 maximum item tests with an IP size of 2 under the FA and Ign conditions. Figure 21C more clearly shows the lower mean NIA conditional on theta for abilities between $\theta = 0.0$ and $\theta = -1.5$ for both the FA and Ign conditions. As IP size is increased to 3 (see Figure 21D), abilities between $\theta = 0.0$ and $\theta = -1.5$ still result in lower mean NIA, although a slight loss of efficiency is lost with slightly higher mean NIAs for these abilities as compared to the IP size of 2 conditions. When IP size is increased to 4 (see Figure 21E), only abilities from $\theta = -1.0$ to $\theta = -0.5$ result in grand mean NIAs below the maximum NIA for the variable length 15 item test conditions under both the FA and Ign conditions.

When test length is increased to 20 items, the range of abilities that have less than the maximum NIA shifts up to $\theta = -1.0$ to $\theta = 0.5$ for the Forced Answer conditions (see Figure 22A). However, the Ignore conditions resulted in a broader range of abilities for which test efficiency is increased, with the range of abilities with less than the maximum NIA ranging from $\theta = -2.0$ to $\theta = 1.0$ (see Figure 22B). Again, as IP size increased, test efficiency decreased, with IP size of 4 resulting in the least efficient tests for the ability levels noted above. Comparing the FA and Ign conditions for the IP size of 2 variable length 20 maximum item test conditions (see

Figure 22C), the increased efficiency for abilities between $\theta = -2.0$ and $\theta = 1.0$ can be seen for both the FA and Ign conditions, with the most efficient tests at $\theta = -0.5$. Increasing IP size to 3 (see Figure 22D) resulted in a slight loss in efficiency for abilities between $\theta = -2.0$ and $\theta = 1.0$, with slightly higher mean NIAs for these abilities. This pattern continues for IP size of 4 (see Figure 22E), with a loss of efficiency for those abilities between $\theta = -2.0$ and $\theta = 1.0$; however, the most efficient tests were those for $\theta = -0.5$ in both FA and Ign conditions. As expected, the Ign conditions resulted in more efficient tests than the FA conditions for both the variable length 15 and 20 maximum item test conditions.

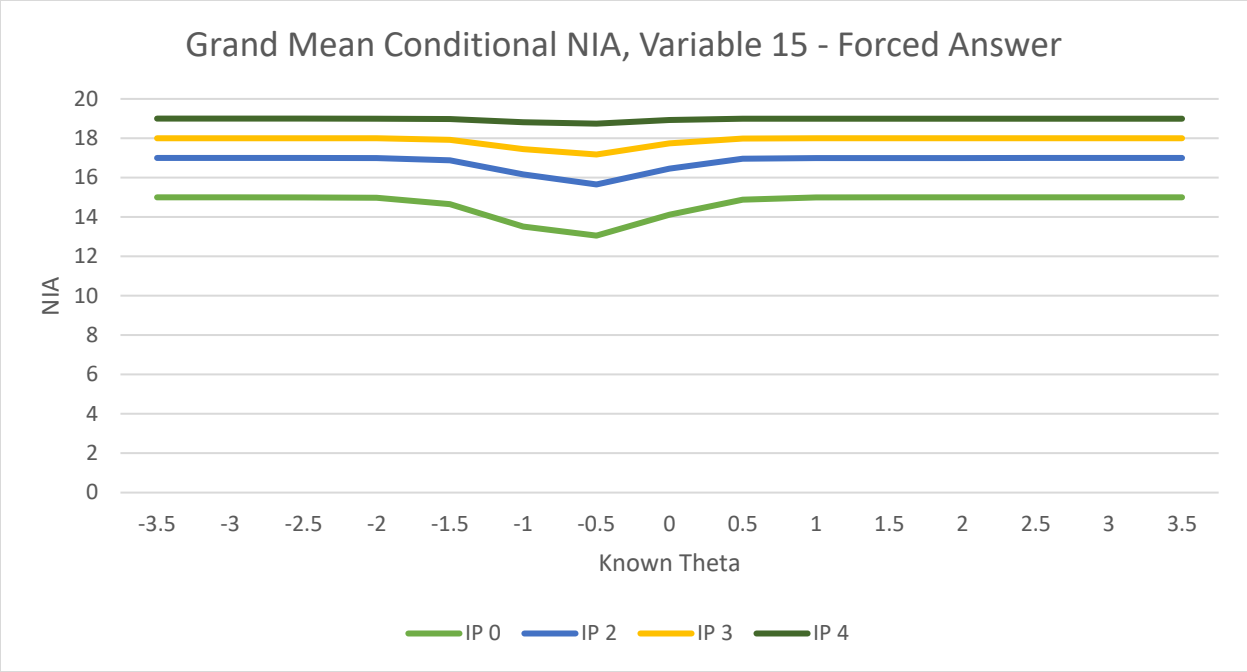

Figure 21A. Grand Mean Number of Items Administered (NIA) Conditional on Known Theta, Variable Length 15 Items, IP Size 0, 2, 3, & 4, Forced Answer Conditions

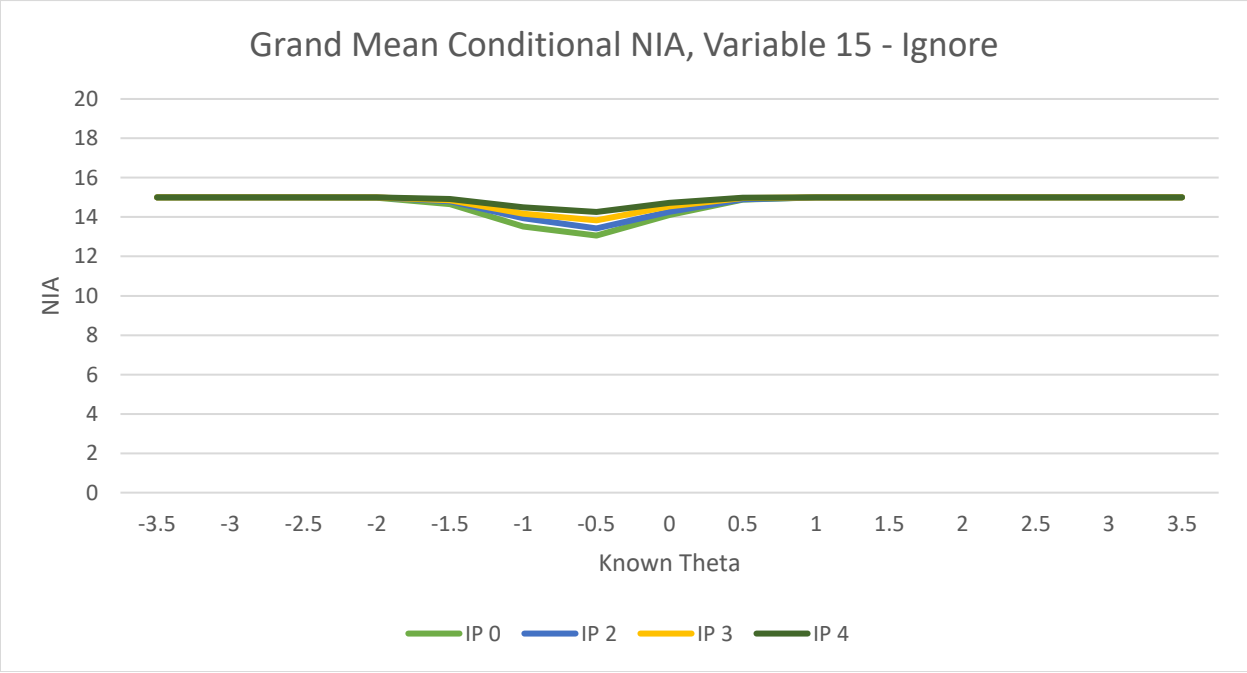

Figure 21B. Grand Mean Number of Items Administered (NIA) Conditional on Known Theta, Variable Length 15 Items, IP Size 0, 2, 3, & 4, Ignore Conditions

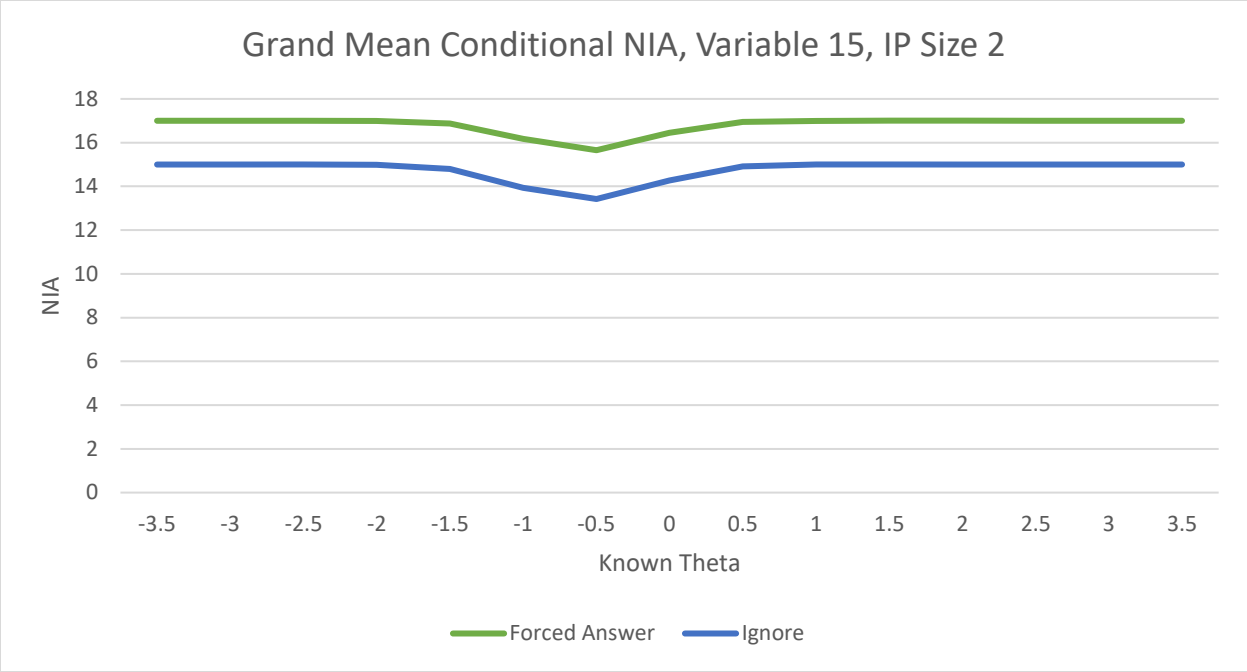

Figure 21C. Grand Mean Number of Items Administered (NIA) Conditional on Known Theta, Variable Length 15 Items, IP Size 2, Forced Answer & Ignore Conditions

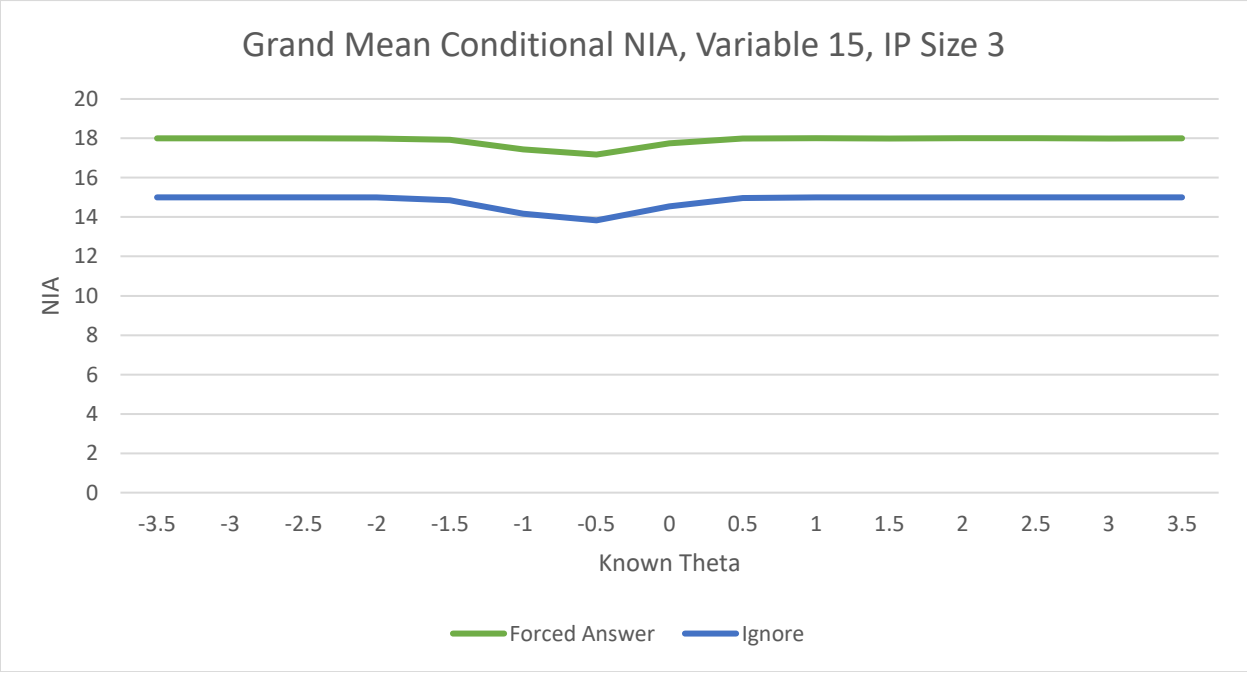

Figure 21D. Grand Mean Number of Items Administered (NIA) Conditional on Known Theta, Variable Length 15 Items, IP Size 3, Forced Answer & Ignore Conditions

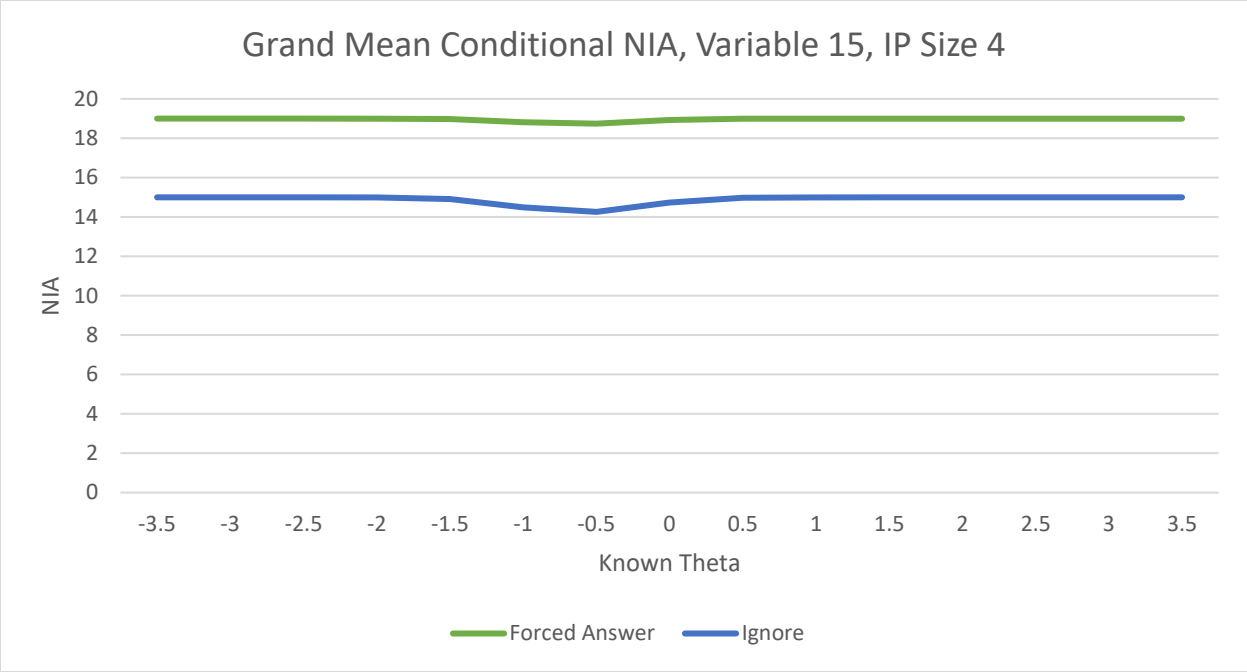

Figure 21E. Grand Mean Number of Items Administered (NIA) Conditional on Known Theta, Variable Length 15 Items, IP Size 4, Forced Answer & Ignore Conditions

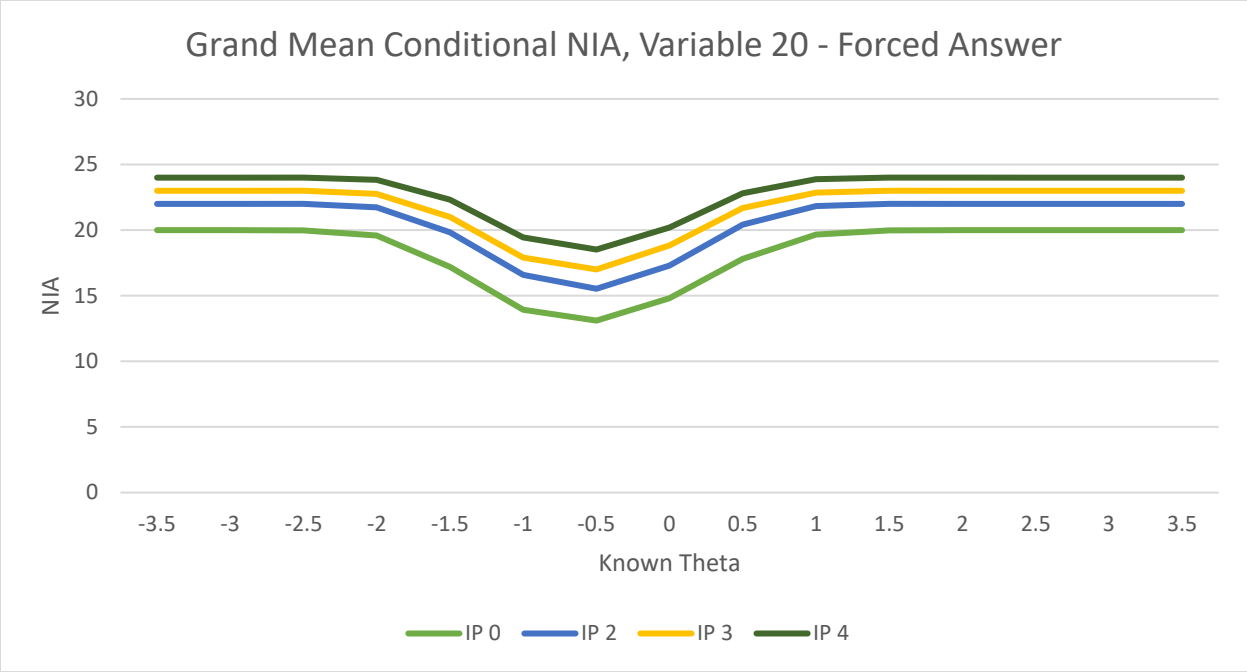

Figure 22A. Grand Mean Number of Items Administered (NIA) Conditional on Known Theta, Variable Length 20 Items, IP Size 0, 2, 3, & 4, Forced Answer Conditions

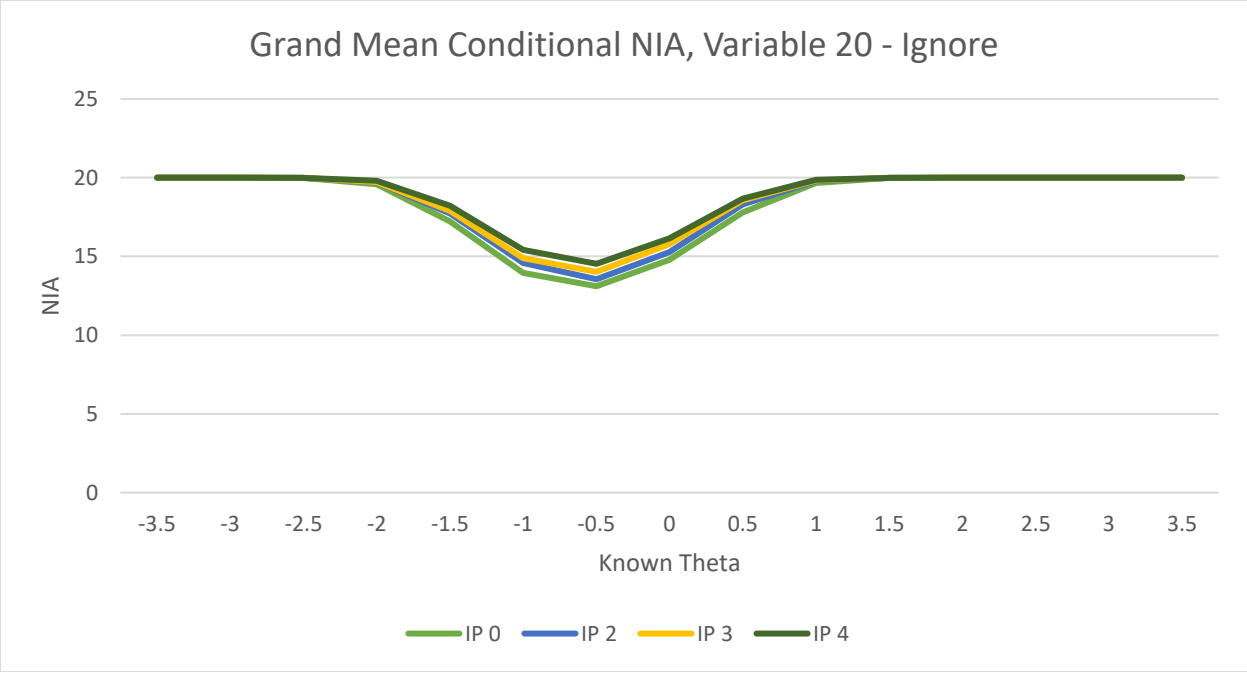

Figure 22B. Grand Mean Number of Items Administered (NIA) Conditional on Known Theta, Variable Length 20 Items, IP Size 0, 2, 3, & 4, Ignore Conditions

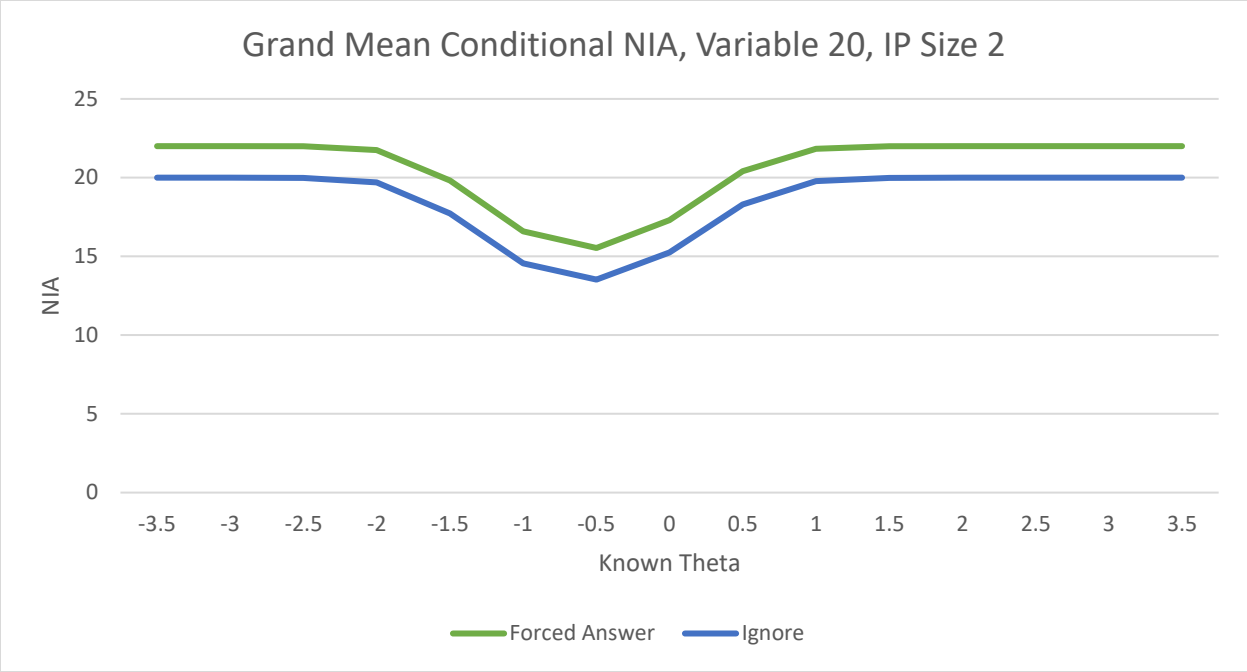

Figure 22C. Grand Mean Number of Items Administered (NIA) Conditional on Known Theta, Variable Length 20 Items, IP Size 2, Forced Answer & Ignore Conditions

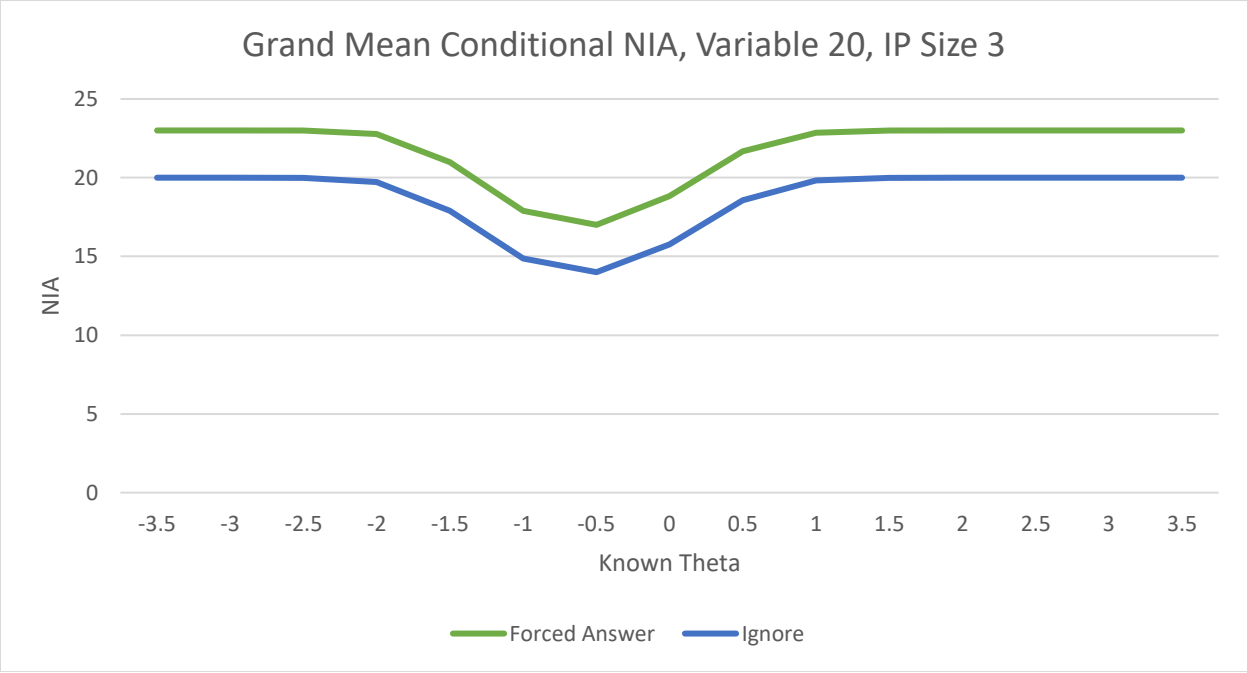

Figure 22D. Grand Mean Number of Items Administered (NIA) Conditional on Known Theta, Variable Length 20 Items, IP Size 3, Forced Answer & Ignore Conditions

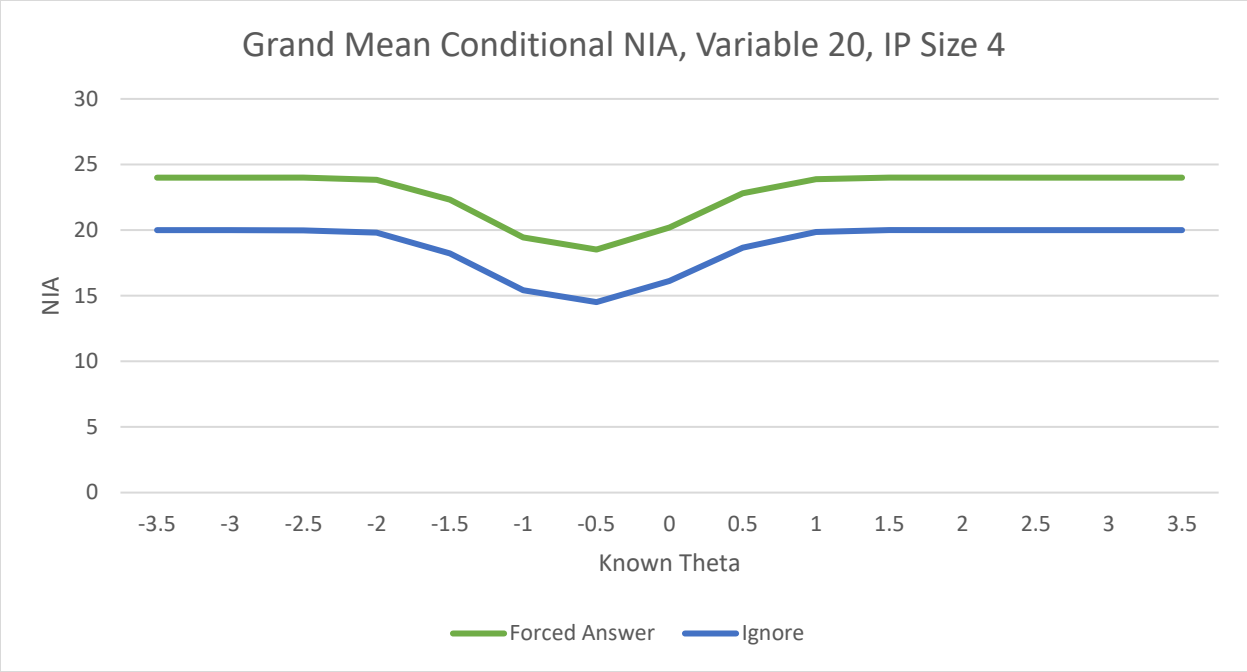

Figure 22E. Grand Mean Number of Items Administered (NIA) Conditional on Known Theta, Variable Length 20 Items, IP Size 4, Forced Answer & Ignore Conditions

Chapter 5: Discussion

Computerized adaptive testing (CAT) has provided the benefit of shorter tests while increasing measurement precision, which is appealing to both educators and test takers. These advances in testing have come with some restrictions. Specifically, the restriction on response review and revision that is allowed with paper-and-pencil (P & P) tests. Recently, a new method, the Item Pocket method, was developed with the intent of relaxing these restrictions while maintaining the benefits of shorter tests and higher measurement precision. The purpose of this dissertation research was to extend the application of this new method to a CAT using a polytomous Item Response Theory (IRT) model that is appropriate for partial credit scoring of items. In addition, the impact of implementation of the IP method on variable length tests was also investigated. A simulation study was conducted, manipulating IP size, test length, termination criteria, and item completion for items left in the IP at the end of the test.

Three main sections in this chapter discuss the study results. The first section outlines the research questions based on the study findings. The next section addresses the limitations of the study and the potential directions for future research. The last section addresses the educational importance and practical applications of the study findings and conclusions are discussed.

Research Questions

1. *What is the impact of the IP method on precision of measurement, across the range of ability levels, when applied to a CAT using the GPCM with Content Balancing and Exposure Control procedures?*

The overall precision of measurement was assessed in multiple ways, one of which was the recovery of the known thetas. The known thetas were generated with a mean of 0 and a

standard deviation of 1.0. The grand mean and standard deviation of the final theta estimates across the 500 replications in each condition were used as descriptive measures of recovery of known thetas. All conditions resulted in grand means and standard deviations very close to a mean of 0 and a standard deviation of 1.0. More specifically, the grand mean and standard deviations of the theta estimates across ability levels in the IP conditions resulted in slightly larger values compared to those in the traditional CAT conditions. The grand means and standard deviations for the 20 item test length conditions resulted in lower values compared to the shorter 15 item test length termination criteria, which was expected due to the increase in the number of items administered. Although the values were larger in all IP sizes as compared to the traditional CAT conditions, the average difference of 0.02 was not practically important.

The mean standard errors (SEs) in each condition across replications also assessed the precision of measurement, with lower SEs indicating more precise measurement. Again, the longer tests resulted in more precise measurement with lower mean SEs in all of the conditions. The mean SE for all IP size conditions was comparable to that of the traditional conditions. The IP size of 4 conditions resulted in the lowest mean SEs in the Forced Answer conditions as a result of the administration of more items. In general, the difference across conditions in mean SEs was less than 0.01, and of no practical importance.

The correlation between the known and estimated thetas was also used to assess precision of measurement. Overall, the IP conditions resulted in comparable mean correlations as the traditional conditions across replications. However, as IP size increased, the mean correlation increased. In the Forced Answer conditions, this is likely due to the administration of additional items, increasing the precision of measurement. The Ignore conditions resulted in equivalent correlations for all of the IP sizes, with slightly larger mean correlations than the traditional

conditions. This indicates that providing an item pocket decreases measurement precision, very slightly, with the average decrease of 0.02 across replications. The differences across IP sizes and the traditional conditions is negligible.

Bias and Root Mean Squared Error (RMSE) were used to assess the precision of measurement, with bias assessing the systematic error in the final theta estimates and RMSE assessing the total error in the final theta estimates. Overall, the mean bias and mean RMSE in the final theta estimates for the IP conditions were similar to those in the traditional conditions across replications. The same pattern as demonstrated with the mean correlations was seen. That is, as IP size increased in the Forced Answer conditions, the mean bias and mean RMSE decreased as compared to the traditional conditions. The Ignore conditions produced identical mean bias and mean RMSE across IP size conditions. This result indicates that the implementation of the IP method results in an average increase in mean bias of -0.01, which is too small to be of any practical importance.

The positive conditional bias seen in Han's (2013) study for the lower abilities was not seen to the same extent in the current study. Han (2013) found that the average bias in theta estimates, within the range of theta from -2 to +2, was 0.057, 0.075, and 0.080 for IP sizes of 2, 4, and 6, respectively (Han, 2013). In the current study, the mean bias was -0.015, -0.013, -0.012, and -0.012 for IP sizes of 0, 2, 3, and 4, respectively, in the fixed length 15 item forced answer test conditions. The fixed length 15 item ignore test conditions resulted in mean bias values of -0.015, -0.016, -0.016, and -0.015 for IP sizes of 0, 2, 3, and 4, respectively. The mean bias decreased slightly with the increase in test length, with the fixed length 20 item forced answer tests resulting in mean bias values of -0.012, -0.012, -0.010, and -0.010 for IP sizes of 0,

2, 3, and 4, respectively. The fixed length 20 item ignore test conditions all resulted in a mean bias of -0.012 for all IP sizes.

Although the mean bias in the current study was calculated using the entire range of theta from -4 to +4, the very slight increase in mean bias with the use of the IP method is an important finding. In addition, the positive conditional bias was not as large in the current study as compared to that found in Han's (2013) study. For instance, the mean conditional bias for $\theta = -2$ was less than 0.05 for the IP size of 0; however, as IP size increased to 2, the mean bias increased to 0.30, and increased slightly more with IP sizes of 4 and 6 to around 0.40 in Han's (2013) study. In the current study, the mean conditional bias for $\theta = -2$ with a fixed length 15 item test was the highest (0.052) with IP size of 4 in the Ign condition and was the lowest (0.039) with IP size of 4 in the FA condition. This mean conditional bias was lower in the longer test conditions, with a high value of 0.036 in the fixed length 20 item test condition with IP size of 0 and a low value of 0.028 on the fixed length 20 item test condition with IP size of 4 with FA. The lack of additional bias in theta estimates is encouraging compared to the significant positive bias found in previous research, which was also more restrictive (Stocking, 1997). Additionally, the substantial decrease in positive conditional bias in the lower ability levels demonstrates the robustness of the IP method to biased ability estimates.

On average, the impact of implementing the IP method appeared to have a very minimal effect on the precision of measurement across the entire range of ability. However, the precision of measurement varied across the ability continuum, requiring the assessment of measurement precision conditional on ability level. The differences seen in the conditional standard errors of measurement (CSEMs) across the range of ability were slight for all conditions compared to the CSEMs in the traditional, IP size of 0 conditions. Generally, the traditional conditions resulted

in slightly higher CSEMs when compared to the other IP sizes in the Forced Answer item completion conditions. This is due to the additional items administered in these conditions, which results in increased measurement precision. Generally, under the Ignore item completion conditions, the CSEMs were slightly higher in the IP size conditions compared to the traditional conditions. This is consistent with the results of Han's (2013) study in which the CSEMs in the IP size conditions resulted in slightly less measurement precision than in the traditional conditions. However, in Han's (2013) study, the CSEMs increased as IP size increased for lower ability levels. This difference in findings can be explained by the additional information provided across a larger range of theta by each polytomously-scored item.

As test length increased, the differences in the CSEMs decreased across the range of ability. The fixed length 20 item tests and the variable length 20 item tests resulted in the most precise measurement, conditional on ability. Smaller differences were seen in the CSEMs between the FA and Ign conditions when IP size increased in fixed length 20 item tests and variable length 20 item test conditions. The forced answer conditions resulted in only slightly more precise measurement in the longer test conditions for both the fixed length and variable length test termination scenarios. Abilities in the middle of the distribution (i.e., $\theta = -1.5$ to $\theta = 0.5$) resulted in the lowest CSEMs, which is due to the item pool distribution. Specifically, because the item pool peaks at $\theta = -0.6$, the item pool as a whole best measures examinees at this ability. As a result of the item pool attributes, the higher abilities generally resulted in slightly larger CSEMs. This pattern was seen in all conditions.

The impact of the IP method on precision of measurement when compared to the traditional conditions is minimal. Recovery of the known thetas, correlations between known and estimated thetas, SE, bias, RMSE, and conditional SEM in IP method conditions were

comparable to those in the traditional CAT conditions without using an item pocket. The inclusion of an IP only slightly decreased measurement precision while creating a more flexible test. These results are consistent with that of Han's (2013) study. The findings suggest that this method could be a viable option to relax restrictions in CATs while maintaining the benefits of CATs.

2. *What is the impact on precision of measurement under the two termination criteria (i.e., fixed and variable length)?*

Han (2013) included only fixed length termination criteria and no items were left in the IP at the end of the test. This study included both fixed and variable length termination criteria, requiring different treatments of the items in the pocket for comparison purposes. Han (2013) forced the administration of the items in the pocket once the examinee approached the maximum number of items allowed for administration. Thus, if an examinee had three items in the pocket and had already been administered 37 items, they would be required to answer the three items in the pocket in order to avoid administering more than the 40 maximum number of items. If this same procedure had been implemented in this study, the two termination criteria would not be comparable. Therefore, the two item completion conditions were included. Forcing the answer of items remaining in the pocket after the termination criteria had been satisfied allows for the comparison of the two stopping rules. However, this also results in more items administered in the Forced Answer conditions than the stopping rule specifies. This has a direct impact on the precision of measurement, resulting in more precise measurement of ability in these conditions.

Overall, more precise measurement occurred in the longer test conditions, both with fixed and variable length tests across replications. On average, the variable length test conditions

resulted in lower grand mean theta estimates, lower correlations between the known and estimated thetas, lower bias, and slightly higher RMSE as compared to the fixed length test conditions. It was expected that fixed length conditions would produce more precise measurement when compared to variable length tests. This is due to the abilities in the center of the ability continuum generally measured more precisely. The fixed length stopping rule will administer items until the maximum number of items has been met, which for abilities in the center of the continuum will result in the lowest SEs for these examinees. The variable length stopping rule terminates the test when the SE drops below the criteria or the maximum number of items has been administered. The abilities in the center of the ability continuum will have SEs at the criteria because the test will stop for them once this achieved, resulting in a mean SE slightly higher than that in the fixed length conditions.

3. *What is the impact of the two of item completion conditions (forced answer or ignored) on precision of measurement?*

It was expected that the item completion conditions would have an impact on the precision of measurement. The Forced Answer conditions resulted in higher precision, as expected, due to the administration of additional items, which results in more precise measurement. In addition, as item pocket size and test length increased, measurement accuracy increased. The Ignore conditions had no effect on precision of measurement across all conditions. When comparing the IP conditions to the traditional conditions, the Forced Answer item completion method resulted in slightly more precise measurement, which increased as IP size increased. The Ignore conditions resulted in slightly less precise measurement, on average, when compared to the traditional conditions, with no differences between IP sizes. In general, the implementation of the IP method in the Ignore item completion condition resulted in a slight

loss of measurement precision. In the Forced Answer conditions, measurement precision is increased, but likely not enough to justify the forced administration of additional items.

4. *What impact does implementation of the IP method have on test efficiency in the variable length conditions?*

Test efficiency was assessed with descriptive statistics of the number of items administered (NIA) and conditional plots of grand mean NIAs. Lower mean values indicate more efficient tests. Overall, the traditional conditions resulted in the lowest mean NIAs across replications. As IP size increased, efficiency decreased slightly, on average. The Forced Answer conditions resulted in the least efficient tests, which is due to the administration of additional items. The Ignore conditions resulted in only slightly less efficient tests as compared to the traditional conditions. The variable length with 15 maximum item tests for all IP sizes in the Ign conditions resulted in only slightly less items administered than the maximum of 15. The variable length 20 item test conditions resulted in 16-17 items administered in the traditional and IP method Ign conditions. This indicates that the IP method does not substantially decrease test efficiency when the items in the pocket are ignored at the end of the test.

The IP method had not been implemented in conjunction with a variable length termination criteria, so a possible interaction was assessed. The mean NIAs across replications could have been affected by the interaction between the IP method and termination criteria. However, due to the sample size of 500,000 simulated examinees, significance would be found for the smallest differences in mean NIAs, if tested. Therefore, the mean NIAs were plotted by IP size in the variable length conditions. Non-parallel lines would be indicative of an interaction. Conversely, parallel lines indicate no interaction. The plots indicated no interaction for all IP

sizes in both variable length 15 and 20 item test conditions. The mean NIAs increased proportionally with IP size in both the Forced Answer and Ignore conditions.

Nonconvergent Cases

The use of maximum likelihood estimation did not occur for cases where either MLE is not reached or ability estimates are out of range (above $\theta = 4.0$ or below $\theta = -4.0$). For these nonconvergent cases, descriptive statistics were first calculated and then these cases were listwise deleted before the outcome measures were calculated. Overall, all of the IP sizes and the two item completion conditions examined resulted in some differences in the mean numbers of out-of-range cases as compared to the traditional CAT without implementing the IP method. However, as IP size increased, the mean number of out-of-range cases decreased, with the traditional CAT conditions resulting in mean numbers of out-of-range cases falling between that of IP sizes 3 and 4. The decrease seen in out-of-range rates with the increase in IP size is likely due to the additional opportunities for MLE to be implemented, with the administration of additional items. In addition, these differences in the mean number of out-of-range cases could be due to the extra information provided by the administration of additional items in the IP Forced Answer conditions. As the IP size increased, the number of items administered increased, which also decreased the number of out-of-range cases. The administration of additional items generally increases opportunities for estimation.

The average number of out-of-range cases in all conditions was similar to the average number of examinees in the extremes of the ability distribution, meaning that the number of out-of-range cases that resulted were expected due to the data generation procedure. The simulated examinees abilities were generated based on a normal distribution with a mean of 0 and a standard deviation of 1, which should result in approximately 2.5% of the abilities in the tails of

the distribution. These abilities would be above a $\theta = 4.0$ and below $\theta = -4.0$, which would result in abilities that are out-of-range.

The inability to estimate an examinee's ability is a serious concern for applied researchers. Therefore, a closer look at the out-of-range cases was completed. Upon further investigation into the out-of-range cases, a limitation of the program used was discovered. The majority of the nonconvergent cases were classified as out of range. However, this was not because those examinees' had ability estimates above $\theta = 4$ or below $\theta = -4$. The variable step size adjustment used to adjust the interim ability estimate before a ML estimate can be obtained did not continue long enough for some of these examinees. This occurred because the high end of the response categories changed depending on the item. For instance, if the response to the first item administered was 4 and that item had 5 response categories, that response would be classified as in one of the extreme categories and the variable step size would adjust the interim ability estimate. If the next response was in the high extreme category, but that item only had four response categories, the variable step size adjustment did not recognize that this was an extreme category and discontinued the adjustment to the interim ability estimate and a standard error was calculated; however, a ML estimate could not be obtained because both responses were in the extreme categories. This resulted in the ability estimate being snapped to a $\theta = 4$, and therefore a nonconvergent case. When the examinee's response was in the lower extreme category for the first few items, the variable step size functioned properly and continued to adjust the interim ability estimate until a response in a middle category was obtained. In addition, if the examinee's first response was in the high extreme response category and the response to the next item was in the low extreme response category, the variable step size adjustment stopped after

the second response when it should have continued until a response that was not in either of the extremes had been received.

The use of content balancing confounded the nonconvergence issue as well. The three content areas contained items with 3, 4, and 5 response categories. For instance, content area one had many more items with 3 response categories compared to the number of items with 5 response categories. For example, there were a total of 6 items in content area three with 5 response categories. This results in an item selected for administration that is not optimal for the examinee's current ability estimate because it may be the only item to select from that content area. When content balancing was not used and the entire item pool as whole was available for item selection, the number of nonconvergent cases decreased by an average of ten cases. It is important to ensure that when content balancing is used that there are a sufficient number of items across the range of abilities within each content area.

The mean number of out-of-range cases decreased as IP size increased, indicating that an interaction between the IP and the ability estimation was present. In order to identify why this was seen, a closer look at the audit trails was conducted. These files contain all of the items selected for administration, the responses, interim ability estimates, and standard errors for each examinee. Inspection of the audit trails revealed that the differential out-of-range cases for the IP size conditions was in part due to the issues seen with the variable step size adjustment and in part due to the way items were selected for placement in the item pocket. The study was a simulation so it was decided that items would be selected for placement in the item pocket if the peak of the item's information function was higher than the examinee's known ability level. If this difference was small, between 0.0 logits and 0.49 logits higher than the known ability, the item would be placed in the item pocket 50% of the time. If this difference was more than 0.5

logits higher than the known ability, then the item would be placed in the item pocket 70% of the time. In the IP size of 2 conditions, if the first item selected was placed in the pocket because it satisfied the criteria for placement in the pocket and it was selected for placement in the IP based on the percentages, either 50% or 70%, the item was placed in the IP and another item was selected for administration. This process continued until either the IP was full or an item was administered. An interesting pattern was noticed. Specifically, when the first two items selected were placed in the IP, the first item actually administered and a response recorded was for the third item selected for administration. The fourth item administered and a response recorded was for one of the first two items placed in the IP. If the responses to the first two items administered were in the high extreme response categories and these were different (i.e., 4 and then 3), the variable step size adjustment stopped and started to calculate the standard error. However, a ML estimate could not be calculated and the interim ability estimate was snapped to the high extreme ($\theta = 4$). The following items selected for administration were selected based on this interim ability estimate, meaning that the items selected were not informative for the examinee's actual ability. This resulted in the inability to recover the known ability.

As the IP size increased, there were more slots for items to be placed if the criteria for placement was satisfied. This criteria was based on the difference between the examinee's known ability and the peak of the item's information function. When the item's information function peak was 0.5 logits higher than the known ability, the item was placed in the IP 70% of the time. When the item's information function peak was closer to the known ability (i.e., 0.0 to 0.49 logits higher than the known ability), the item was placed in the IP 50% of the time. Every time an item was selected for administration, the item was evaluated for placement in the IP, meaning the error that was introduced in placement of items in the item pocket increased with

every additional slot for an item. This was seen in the IP size of 4, with more examinees responding to the first few items rather than the items being placed in the IP. For instance, looking at the same examinee in the fixed length 15 item test with an IP size of 2 condition, the first two items selected for administration were placed in the IP. Thus, the discrepancy between the known ability and the item's information function peak was large enough and the item fell in the 50% or 70% bucket for placement in the pocket. However, in the IP size of 4 condition, the first item selected for administration was administered rather than placed in the IP, meaning that it did not fall into the 50% or 70% bucket for placement in the pocket. The response was then recorded for the examinee in the IP size of 4 condition and the next item was selected for administration based on the updated interim ability estimate. As a result of the first few items selected for administration being administered, the issue with the ML estimation spinning out-of-range decreased and the number of out-of-range cases decreased as well.

Limitations and Future Research

The findings of the current study support the use of the IP method for items that the examinees find challenging. However, due to the simulated nature of the study, items were selected for placement in the pocket based on the difference between the known ability and the peak of the item's information function, indicating the ability level for which the item most precisely measures. When the known ability is further below the item's information function peak, it is assumed that the item would be challenging to the examinee and therefore be placed in the pocket. This may not be the way examinees use the IP in a true testing situation. Thus, the IP method should be studied with live examinees, investigating the true use of the item pocket. However, the current simulation study limited the use of the IP to items the simulated examinee

would find challenging based on the item characteristics compared to the examinees' known ability.

As was the case with Han's (2013) study, there was no time limit for the test. Although in operational tests, most examinees complete the entire test, not all complete the test all the time. The examinees in the lower ability levels are more susceptible to not completing the test. The inclusion of an IP would likely have the effect of further decreasing completion by those examinees in the lower extreme of the ability continuum. However, this was not examined in this study and could limit the applicability in CATs for certain populations of examinees.

Most operational tests are composed of both dichotomous and polytomous items, referred to as mixed format tests. The current study examined the applicability of the IP with polytomously-scored items only. The previous research (Han, 2013) used dichotomously-scored items with the 3-PL model and used the difficulty of the item (the *b*-parameter) for determining which items are placed in the pocket. Polytomous items do not have one *b*-parameter associated with them. Instead, multiple *b*-parameters are associated with polytomously-scored items. Therefore, the item information function peak was used to infer difficulty of the item and placement in the IP. There is the possibility that the findings from Han's (2013) study may slightly change if the 2-PL model was used and item information functions were used for determining use of the IP. The GPCM simplifies to the 2-PL when there are only two score categories. Both the 2-PL and the GPCM could be used in a mixed format test. Therefore, future research should investigate mixed format tests with the IP method. In addition, the IP method should be studied with live examinees, examining the true use of the item pocket. Han suggested that response review and revision restrictions could increase examinee test anxiety, which the IP method could reduce. Thus, it would be interesting to also assess examinee test

anxiety in conjunction with the IP method. Currently, operational tests that allow response review and revision are multi-stage tests, which allow review and revision with each module. Future research could compare these two methods.

A couple of limitations of the current study were discovered through investigation of the nonconvergent cases. The use of maximum likelihood estimation can and does result in nonconvergent cases. There are two types of nonconvergent cases, those in which the final theta estimates are outside of the range of theta, -4 to +4, and those where MLE is not reached. However, every test-taker expects an ability estimate at the conclusion of the test. In the current study, the number of cases where the ability estimate was out-of-range were higher than expected. It was discovered that the variable step size adjustment was not performing as designed, which resulted in increased out-of-range cases. The use of Bayesian estimation would avoid this issue, such as Expected a Posteriori (EAP) estimation (Bock & Mislevy, 1982), which is most commonly used in CAT. In addition, the limited number of items in some of the content areas confounded the out-of-range issue. It is recommended that the item pool contain a sufficient number of items in all content areas. In the current study, the number of items in the content area with four or five response categories was very low and had a direct impact on nonconvergence. The process of selecting items for placement in the item pocket was a limitation to the current study as well. The IP appeared to interact with the issues discovered with the variable step size adjustment. As the IP size increased, this issue decreased, resulting in less out-of-range cases. Consequently, the simulation program used in the present study should be modified to allow for the appropriate administration of items and estimation of ability using MLE in the scenarios wherein ability estimates would be estimated as out-of-range as previously described. Further, the selection of items for placement in the IP may not be the actual way real

examinees will use the IP. Therefore, further research is needed studying the true use of the IP with live examinees.

Educational Importance

The use of Computerized Adaptive Testing has increased in the last few decades with the increase of computer use in daily life. However, the majority of tests that students are exposed to are still of the paper-and-pencil (P & P) variety. CATs are typically more restrictive than P & P tests due to the algorithms used. Adaptive testing has benefits of shorter tests and more precise measurement of abilities. However, there is a lack of flexibility with CATs to be able to move through the test like a P & P test and review and/or revise responses. Previous research (Vispoel et al., 2000) found that examinees desire the ability to review and revise answers in CATs. Until the use of the IP method, review and revision of responses with CATs was restricted to occur after all the items had been answered, resulting in biased ability estimates. The IP method provides the opportunity to relax these restrictions while maintaining acceptable measurement precision.

The findings from the initial IP method study (Han, 2013) indicated that the IP method can be applied to a CAT using the 3-PL model and maintain an acceptable level of measurement precision. The current study extended this line of research to a polytomous IRT model appropriate for partial credit scoring of items. The findings indicate that measurement precision is comparable to that of a CAT without using the IP method. In addition, the performance of the IP method in conjunction with a variable length termination criteria was explored. Results indicated that implementation of the IP method with a variable length test produces comparable measurement precision and test efficiency to a CAT without implementing the IP method.

The optimistic findings of comparable measurement precision to a traditional CAT and reduced bias as compared to Han's (2013) findings support the use of the IP method with CATs using partial credit scoring of items. Specifically, IP sizes of 2, 3, and 4 for the fixed length 20 item FA test conditions resulted in the higher mean correlations, which increased as IP size increased as compared to the traditional condition, indicating that the FA improves recovery of known thetas. The fixed length 20 item test conditions also resulted in lower mean bias and RMSE as compared to the traditional condition, which decreased as IP size increased in the FA conditions. The mean final theta estimates were closer to 0 and smaller mean SEs in the fixed length 20 item FA test conditions as compared to the traditional and the shorter 15 item fixed length FA test conditions. The measurement precision increased as IP size increased. This increased measurement precision comes at the cost of test efficiency with mean NIAs higher than those seen in the traditional and in the Ign conditions. The fixed length 15 and 20 item Ign test conditions resulted in slightly less measurement precision, but also resulted in lower mean NIAs than the FA test conditions. These findings are consistent with Han's (2013) study, while extending the applicably to fixed length tests using polytomously-scored items appropriate for partial credit scoring of items.

In addition, the findings support the applicability of the IP method to variable length tests, which account for a sizable portion of operational CATs. The longer variable length 20 item FA test conditions resulted in the highest measurement precision, which increased as the IP size increased as compared to the traditional variable length 20 maximum item test conditions and the corresponding Ign conditions. Again, this increased measurement precision decreased test efficiency in the FA variable length test conditions.

The differences seen in the conditions are minimal and produce practically equivalent measurement precision. Therefore, the tester should choose the type of test to administer (e.g., traditional CAT versus an IP CAT) based upon their needs. The size of the IP should be based on the test length, restricting the size of the pocket to hold only 20% of the items to be administered. The decision of how to handle the items left in the IP at the end of the test should also be determined by the tester. Ignoring the items could threaten over-exposure of those items; however, if exposure is not a concern, then ignoring the items in the pocket results in practically equivalent measurement precision as a traditional CAT. Overall, this method allows CATs to be less restrictive and more like P & P tests, which students are used to taking. Additionally, this research expands the types of CATs to which the IP method could be applied. The application of this method in a live testing situation could possibly reduce examinee anxiety due to the relaxed restrictions and possibly reduce careless examinee errors, although little to no research exists on anxiety due to CAT restrictions.

REFERENCES

Andrich, D. (1978). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement*, *2*, 581-594.

Birnbaum, A. (1958). *On the estimation of mental ability.* Series Report No. 15. Project No. 7755-23, USAF School of Aviation Medicine, Randolph Air Force Base, Texas: January.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Bock, D. R. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29–51.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*, 431-444.

Boyd, A. M. (2004). Strategies for controlling testlet exposure rates in computerized adaptive testing systems [Doctoral dissertation, University of Texas at Austin, 2003]. *Dissertation Abstracts International, 64,* 11, 5835B (No. AAT 3110732).

Boyd, A. M., Dodd, B. G., & Choi, S. W. (2010). Polytomous models in computer adaptive testing. In *Development and applications of polytomous item response.* Philadelphia, PA: Francis Taylor.

Chang, H.-H., Qian, J., & Ying, Z. (2001). a-Stratified multistage computerized adaptive testing with b blocking. *Applied Psychological Measurement*, *25*(4), 333-341.

Chang, H.-H., & van der Linden, W. J. (2003). Optimal stratification of item pools in {alpha}-Stratified computerized adaptive testing. *Applied Psychological Measurement*, *27*, 262–274.

Chang, H.-H., & Ying, Z. (1999). a-Stratified multistage computerized adaptive testing. *Applied Psychological Measurement*. *23*, 211-222.

Chen, S.-K., Hou, L., & Dodd, B. G. (1998). A comparison of maximum likelihood estimation and expected a posteriori estimation in CAT using the partial credit model. *Educational and Psychological Measurement*, *58*(4), 569-595.

Choi, S. W., & Swartz, R. J. (2009). Comparison of CAT item selection criteria for polytomous items. *Applied Psychological Measurement, 33*(6), 419-440.

Davis, L. L. (2004). Strategies for controlling item exposure in computerized adaptive testing with the generalized partial credit model. *Applied Psychological Measurement, 28,* 165-185.

Davis, L. L., & Dodd, B. G. (2003). Item exposure constraints for testlets in the verbal reasoning section of the MCAT. *Applied Psychological Measurement, 27,* 335-356.

Dodd, B. G., & De Ayala, R. J. (1994). Item information as a function of threshold values in the rating scale model. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (Vol. 2, pp. 301-317). Norwood, NJ: Ablex.

Dodd, B. G., De Ayala, R. J., & Koch, W. R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement*, *19*, 5-22.

Dodd, B. G., Koch, W. R., & De Ayala, R. J. (1993). Computerized adaptive testing using the partial credit model: Effects of item pool characteristics and different stopping rules. *Educational and Psychological Measurement*, *53*(1), 61–77.

Dodd, B. G., & Koch, W. R. (1987). Effects of variations in item step values on item and test information in the partial credit model. *Applied Psychological Measurement*, *11*, 371-384.

Embertson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Gorin, J. S., Dodd, B. G., Fitzpatrick, S. J., & Shieh, Y. Y. (2005). Computerized adaptive testing with the partial credit model: Estimation procedures, population distributions and item pool characteristics. *Applied Psychological Measurement, 29*(6), 433-456.

Han, K. T. (2013). Item Pocket Method to Allow Response Review and Change in Computer Adaptive Testing. *Applied Psychological Measurement*, *37*(4), 259–275.

Kamakura, W., & Srivastava, R. R. (1982). Latent Trait Theory and Attitude Scaling: the Use of Information Functions For Item Selection. *Advances in Consumer Research, 9*, *eds. Andrew Mitchell, Ann Abor, MI: Association for Consumer Research,* 251-256.

Kingsbury, G. (1996). *Item review and adaptive testing.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, *2*, 359-375.

Koch, W. R., & Dodd, B. G. (1989). An investigation of procedures for computerized adaptive testing using partial credit scoring. *Applied Measurement in Education, 2,* 335-357.

Lee, Y. H., Ip, E. H., & Fuh, C. D. (2007). A strategy for controlling item exposure in multidimensional computerized adaptive testing. *Educational and Psychological Measurement*, *68*(2), 215–232.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Lunz, M. E., Bergstrom, B. A., & Wright, B. D. (1992). The effect of review on student ability and test efficiency for computerized adaptive tests. *Applied Psychological Measurement*, *16,* 41-51.

Lunz, M. E., & Stahl, J. A. (1998). *Patterns of item exposure using a randomized CAT algorithm.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–173.

McClarty, K. L., Sperling, R. A., & Dodd, B. G. (2006, April). *A variant of the progressive-restricted item exposure control procedure in computer adaptive testing systems based on the 3PL & partial credit models,* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement, 13,* 57-75.

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement*. *16*, 59-71.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159-176.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70,* 351-356.

Parshall, C., Harmes, J.C., & Kromrey, J. D. (2000). Item exposure control in computer-adaptive testing: The use of freezing to augment stratification. *Florida Journal of Educational Research, 40*(1), 28-52.

Pastor, D. A., Dodd, B. G., & Chang, H.-H. (2002). A comparison of item selection techniques and exposure control mechanisms in CATs using the generalized partial credit model. *Applied Psychological Measurement*, *26*, 147-163.

Penfield, R. D. (2006). Applying Bayesian item selection approaches to adaptive tests using polytomous items. *Applied Measurement in Education, 19*(1), 1-20.

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

Reckase, M. D. (1989). Adaptive testing: The evolution of a good idea. *Educational Measurement Issues and Practice*, *8*, 11–15.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, *35*, 311–327.

Rost, J. (1988). Measuring attitudes with a threshold model drawing on a traditional scaling concept. *Applied Psychological Measurement*, *12*, 397-409.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, Monograph 17.

Stocking, M. L. (1997). Revising item responses in computerized adaptive tests: A comparison of three models. *Applied Psychological Measurement*, *21*(2), 129–142.

Stone, G. E., & Lunz, M. E. (1994). The effect of review on the psychometric characteristics of computerized adaptive tests. *Applied Measurement in Education*, *7*, 211-222.

Sympson, J. B., & Hetter, R. D. (1985, October). *Controlling item exposure rates in computerized adaptive testing,* Paper presented at the annual meeting of the Military Testing Association, San Diego.

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer (Ed.), *Computer adaptive testing: A primer* (2nd ed., pp. 101-133). Hillsdale, NJ: Lawrence Erlbaum.

van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika, 63*(2), 201-216.

Vispoel, W. P., Clough, S. J., Bleiler, T., Hendrickson, A. B., & Ihrig, D. (2002). Can examinees use judgments of item difficulty to improve proficiency estimates on computerized adaptive vocabulary tests? *Journal of Educational Measurement*, *39*, 311–330.

Vispoel, W. P., Hendrickson, A. B., & Bleiler, T. (2000). Limiting answer review and change on computerized adaptive vocabulary tests: Psychometric and attitudinal Results. *Journal of Educational Measurement*, *37*, 21–38.

Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice*, *15*, 15–20.

Wainer, H. (2010). *Computerized adaptive testing: A primer* (2nd Ed.). Mahwah, NJ: Routledge.

Wainer, H., & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics*, *12,* 339-368.

Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice*, *17*(4), 17–27.

Whittaker, T. A., Fitzpatrick, S. J., Williams, N. J., & Dodd, B. G. (2003). IRTGEN: A SAS macro program to generate known trait scores and item responses for commonly used item response theory models. *Applied Psychological Measurement*, *30*, 299-300.

Wise, S., Finney, S., Enders, C., Freeman, S., & Severance, D (1999). Examinee judgments of changes in item difficulty: Implications for item review in computerized adaptive testing. *Applied Measurement in Education, 12,* 185-198.