IMPACT OF COMPOSITE POPULATION PRIORS ON COMPUTER ADAPTIVE TEST PROFICIENCY ESTIMATES

A Dissertation
Presented to
The Academic Faculty

By

Kristin M. Morrison

In Partial Fulfillment
Of the Requirements for the Degree
Doctor of Philosophy in Psychology

Georgia Institute of Technology
August 2017

IMPACT OF COMPOSITE POPULATION PRIORS ON COMPUTER ADAPTIVE TEST PROFICIENCY ESTIMATES

Approved by:

Dr. Susan E. Embretson, Advisor
School of Psychology
Georgia Institute of Technology

Dr. Rick Thomas
School of Psychology
Georgia Institute of Technology

Dr. James S. Roberts
School of Psychology
Georgia Institute of Technology

Dr. Jonathan Templin
Department of Educational Psychology
University of Kansas

Dr. Daniel Spieler
School of Psychology
Georgia Institute of Technology

Date Approved: 10 April 2017

ACKNOWLEDGEMENTS

First, I would like to thank Dr. Susan Embretson for her guidance, expertise, support, and patience as I navigated my way through this process known as grad school. I am grateful to Dr. James Roberts, who was always willing to lend his guidance and support when I needed them. I would like to thank my dissertation committee, Dr. Dan Spieler, Dr. Jonathan Templin, and Dr. Rick Thomas, for their insightful and constructive feedback. My sincerest gratitude goes out to Megan Lutz for her friendship, her supportiveness to remind me I could succeed, and her willingness to illuminate the path with her insights when I was stranded in the dark. My thanks also go out to all my friends who have provided love and support along the way. I would not have accomplished all that I have without their kind words.

I would never have achieved this accomplishment without the love of my *little* brother, who, although he is many years younger, is wise beyond his years. He always knew how to support and push me to achieve my goals, and was always there to remind me just how awesome I could be. Last, but not least, my eternal thanks go to my parents, who are quite literally my biggest fans. They always told me to shoot for the stars and believed that I could attain any goal I set my sights on. Without their unconditional love and support, I would never have set goals of this magnitude for myself. I hope to continue to make them proud on the next step of my journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Testing has existed for thousands of years and has evolved from all examinees receiving the same test to adaptive testing, in which the test is tailored to the individual examinee. These adaptive testing designs have shown to be improvements over fixed-length, conventional tests in terms of proficiency measurement, reduced testing time, and faster scoring. However, these designs introduce a variety of issues that must be considered during test development as well as during the test's lifespan. A popular method of test administration is computer adaptive testing (CAT) using *expected a posteriori* (EAP) estimation. This Bayesian estimation approach utilizes previous information known about the examinee to obtain more precise estimates of the individual's ability. An appropriate prior will generally increase estimation precision, decrease outlier influences, and provide an estimate for all possible response patterns. An inappropriate prior, however, may result in biased estimates (Embretson & Reise, 2000). Previous studies have used collateral information (i.e., additional information) concerning the examinee to aid in estimation. This collateral information may be related to item properties, such as item difficulty, or related to the individual, such as demographic variables, age, grade, or previous test scores. Several studies have used previous test scores (Matteucci & Veldkamp, 2013; Veldkamp & Matteucci, 2013; van der Linden, 1999) as collateral information, none have looked directly at priors based on group membership. This study examined the influence of various group priors, such as composite priors (i.e., priors created from combining groups) and individual priors (i.e., priors specific to the group), on estimation in CAT designs. Results of the study show group-specific priors perform best; however, it is impossible to know an individual's true group. Thus, results of the study support the use of priors based on the population because priors based on demographics may adversely impact some high ability groups.

CHAPTER 1

INTRODUCTION

In psychology, an individual's behavior repertoire is used to infer his or her latent traits, or abilities, which are relatively stable attributes (e.g., skills or abilities) of the individual that cannot be directly measured (θ; Crocker & Algina, 2008; Green, Bock, Humphreys, Linn, & Reckase, 1984). For example, height and weight are directly measureable aspects of an individual. However, an individual's social proclivity, verbal ability, or mathematics ability are not directly measureable and are considered latent constructs. Their existence cannot be determined absolutely and must be inferred by examining behavior. Once the existence of a construct has been proposed and a sufficient definition linking the construct to observable behaviors has been provided, instruments can be developed that set forth a systematic procedure (i.e., a test) for obtaining behavior samples (Crocker & Algina, 2008). For example, an educational psychologist may be interested in a student's mathematical achievement, a construct that has been established to exist but is not directly observable. The psychologist would first need to specify a link between mathematics ability and behaviors that can be observed from the student. These behaviors might include the number of mathematics items, utilizing addition, subtraction, multiplication, and division, the student is able to answer correctly in a pre-determined time frame. The test is the student's behavior on these items. Thus, the psychologist can infer the level of mathematics ability the student possesses by assigning a quantitative value (e.g., number of items answered correctly) to the behavior; this is called measurement (Crocker & Algina, 2008).

Tests can be administered in many different formats. One test administration aspect deals with administration mode. A test may be administered as a paper-based test (PBT) or a

computer-based test (CBT). CBTs are administered using computer technologies and have been shown to have many improvements over PBTs. However, more decisions must be made when using a CBT, such as item development, test assembly, test composition, examinee issues, test delivery, and post-test procedures (Luecht, 2006). Another aspect of tests deals with adaptability. Tests may have no adaptation and all examinees receive the same set of items; these are termed conventional tests (CTs). In this testing design, examinees may be presented with items that are too difficult or too easy, and thus the items provide little information concerning the examinee pertaining to the construct of interest (Yan, Lewis, & von Davier, 2014). Or, a test may be adaptive, in which different individuals receive different sets of items based on their responses to previous items. These tests are optimally designed, as examinees are presented with items that are ideal for their ability and thus are neither too difficult nor too easy (Meijer & Nering, 1999; Weiss, 1985). Adaptive tests are often superior to conventional tests for their efficient and precise measurement of the examinee's ability (Weiss, 1985; Yan et al., 2014).

Although linear CBTs (e.g., a conventional test administered via a computer) can be administered, capitalization on the efficiency of adaptive testing is often married with the advantages of computer-based administration. Through this marriage and the use of item response theory (IRT), a sophisticated approach that uses both person and item characteristics to measure latent traits (Embretson & Reise, 2000), two types of adaptive testing have been developed. These approaches to automated, adaptive testing are computerized adaptive testing (CAT) and multistage testing (MST).

CATs are adaptive assessments in which the test is tailored to the examinee at the individual item-level. As the test progresses, the adaptive algorithm hones in on the examinee's ability level by using responses to all previously administered items, administering new items

that provide the most information about the individual (Meijer & Nering, 1999; Weiss, 1985; Weiss & Kingsbury, 1984). Compared to conventional tests, it is possible that every examinee will receive a different set of items (i.e., a different test). This approach to testing has many advantages over conventional tests. CATs often result in more precise trait measurement, require less items, involve shorter amounts of time to administer, can be scored immediately, and thus, can result in immediate score reporting to the examinee. Although CATs have many advantages, some disadvantages do exist, such as starting costs, best ways to make classification decisions concerning examinees, how to control content, and a more complicated item review process.

MSTs, also called computer adaptive sequential tests (CASTs), are also adaptive assessments that tailor the test to the examinee's ability, but instead of adapting at the item-level, these tests adapt at the item-set level (Yan et al., 2014). In MSTs, groups of items are administered to examinees based on previous responses to item groups. These tests can be considered a special case of CATs and aim to capitalize on CAT advantages, such as more precise measurement, while minimizing their disadvantages. Thus, MSTs allow for greater control of content, since the item groups can be scrutinized before administration, and also allow examinees to review their responses to items within an item group before proceeding with the test (Mead, 2006).

Often, the goal of a test, whether PBT, CBT, conventional, or adaptive, is to accurately measure an individual's latent ability or to classify individuals into various groups (Weiss & Kingsbury, 1984; Yan et al., 2014). These goals are often determined by the test's purpose. For example, in an educational setting, the goal of a test may be to obtain an accurate estimate of a

student's mathematics ability to rank order students. However, in this same setting, the goal might be to classify a student as proficient or not proficient (i.e., classification).

Multiple IRT approaches exist to obtain an examinee's position on the latent trait continuum. IRT uses the behavior of the examinee, their scored responses on the items, to achieve this estimate. For example, for dichotomously scored items, a response pattern can be obtained for each person. Thus, examinees *A* and *B* may be administered the same set of items but have different response patterns (*RP*). Examinee *A* may answer the first five questions correctly and the last five questions incorrectly $\{RP_A = 1111100000\}$, while Examinee *B* might have a different pattern $\{RP_B = 0111011000\}$ but still answered five questions correctly. Classical test theory (CTT) would give both of these examinees the same number-right score (i.e., 5). However, each examinee endorsed different items (i.e., answered different items correctly) and might vary on the latent trait. IRT trait estimation is equipped to handle this via likelihood estimators (e.g., Maximum Likelihood Estimation) or Bayesian estimators (e.g., Expected A Posteriori Estimation). When estimating ability, likelihood estimates are often less biased but Bayesian estimators are often more precise.

Bayesian estimators have a unique quality – they allow the introduction of a person prior distribution in the estimation process. The prior distribution is a hypothetical distribution from which the examinees are a random sample. This prior gives more examinee information, increasing both the efficiency of the test and the precision of the ability estimate, will protect against the influence of outliers, and can provide an estimate for all possible response patterns. However, if an inappropriate prior distribution is chosen, the resulting ability estimates may be biased, especially for extreme abilities (Embretson & Reise, 2000). A common prior distribution

is the standard normal distribution, which has a mean of 0 and a standard deviation of 1 ($\theta \sim N(0, 1)$). However, any appropriate distribution can be used.

Several techniques are used to ensure a test is unbiased and fair in terms of different population subgroups; however, often group differences in trait levels for various cognitive abilities do exist. Aspects of cognitive ability (e.g., general intelligence, spatial ability, memory, etc.) are measured via different testing techniques, since different patterns of ability can exist. For example, two demographically similar individuals might be compared on their cognitive abilities. One individual might be high in general intelligence, low in verbal ability, and high in mathematical ability. The other individual might be high in all areas. The different aptitudes are measured since not all abilities are correlated with demographic information, and although relationships may exist between abilities, these relationships are not always consistent. Differences in test performance have been documented between Caucasians, African Americans, and Hispanics in performance on various educational, military, and personnel selection assessments (Roth, Bevier, Bobko, Switzer, & Tyler, 2001). For example, overall differences in general intelligence are typically found; Caucasians scored higher than the other two ethnicities. The Graduate Record Examination (GRE; ETS, 2016) and the SAT (SAT, 2015) report gender and ethnic differences in means and standard deviations. For the GRE, females tend to score lower than males, and African Americans score lower than Caucasians.

Collateral information pertaining to the examinees can be used to specify the prior distribution. This information may be demographic variables (e.g., gender, ethnicity), socioeconomic status, grade, country, age, or previous test scores. Using this information, different priors may be utilized for trait estimation based on the examinee's status on the covariate (e.g., male versus female). Studies have utilized empirical priors for ability estimation

by creating relationships between ability and the collateral information. For instance, Veldkamp and Matteucci (2013) used performance on one construct as collateral information to create an informed, empirical prior for performance on a similar construct. These approaches to ability estimation provide better provisional estimates in adaptive testing, which might decrease the time needed to converge on the true ability (i.e., shorter tests), as well as increased statistical precision and lower item exposure (Matteucci & Veldkamp, 2013; van der Linden, 1999; Veldkamp & Matteucci, 2013).

Motivation and Research Questions

While collateral information, in the form of previous test scores, has been used in studies concerning ability estimation, no study has directly examined the relationship between ability estimation and collateral information based on group membership. Therefore, this study examines the impact on ability estimation when different group priors are utilized, and will inspect the influence of these different priors on estimation both between (i.e., Group A versus Group B) and within (i.e., high versus low ability examinees) groups. As stated, most Bayesian estimation applications utilize a standard normal prior for all examinees. However, different prior distributions based on collateral information (i.e., group membership) may exist. For example, Group A might have a lower mean prior distribution [$\theta_A \sim N(-0.5, 1)$] while Group B has a higher mean prior distribution [$\theta_B \sim N(0.5, 1)$]. While both of these distributions are normal distributions, they vary in their mean ability. Thus, these individual priors could potentially be used to estimate ability but may differentially influence various aspects of ability estimation and its use. For example, while Group A might have a lower population mean, high ability examinees in this group may be negatively impacted when this prior is used versus a standard normal prior.

The simulation will vary the test type (i.e., conventional, CAT), the true ability distribution of the examinees, prior distributions used during estimation based on collateral information, final ability estimation approaches, and IRT model. These various testing conditions will be examined for measurement precision via the standard error of the ability estimate, bias, as well as accuracy via Root Mean Square Error (RMSE). In general, the simulation will aim to examine the impact on trait estimation for individuals within and across various groups when prior distributions are chosen based on group membership.

Focus of the current study is aimed towards educational assessments, such as high-stakes tests administered to students. While the results may support the use of group priors, other variables will need to be considered before the method is applied. For example, if results of the simulation support the incorporation of group priors into testing designs, testing companies may need to examine any legal ramifications of the approach beyond the psychometric ones. The approach may not be advantageous for a testing company if it opens up legal avenues for the company to be sued.

However, even though the approach may not be advantageous in education settings, it may have applicability in other domains. One possible domain is health screenings. People may be partitioned into groups based on answers to questions, and each of these groups may have a different probability of a health concern. For example, an individual may have a higher risk of a type of cancer. By putting the individuals into groups based off various factors (i.e., covariates), risk values can be calculated. Another possible domain is in personnel selection and job placement.

CHAPTER 2

LITERATURE REVIEW

This chapter contains a brief review of the literature. Topics will include an examination of item response theory, which will encompass explanations concerning various models as well as approaches to trait estimation and classification. Different testing designs will be explained and compared. A summary of subgroup differences relating to performance will be given, and the chapter will conclude with approaches to utilizing collateral information (i.e., information about the examinee) during testing.

Item Response Theory

Item response theory (IRT), also known as latent trait theory, has become the predominant approach to psychological measurement (Embretson & Reise, 2000). Measurement of performance in IRT depends on the relationship between the characteristics of items administered to an individual and the individual's responses to those items. Thus, a relationship exists between the individual's observable item performance and the underlying, unobservable trait (θ) the items aim to measure. Lord and Novick (1968) expressed the need to understand this relationship as a move to individualized testing became possible. Item response models specify the specific relationship between the observable and unobservable variables and are considered "strong models" because of the stringent assumptions placed on the data that are not easily met (Hambleton & Jones, 1993).

Three assumptions underlie IRT models. The first assumption is the dimensionality assumption, which states that a specific number of dominant latent variables underlyie behavior on the observed variables (de Ayala, 2008; Hambleton, Swaminathan, & Rogers, 1991). A single dominant trait (e.g., mathematics ability) may underlie examinee behavior. When one

dominant trait exists, unidimensional IRT models are appropriate to use. However, it is possible that multiple dominant traits (e.g., mathematics and reading ability) may be necessary in order to sufficiently explain behavior and performance (Hambleton et al., 1991). In this case, multidimensional IRT (MIRT) models are used to describe the interaction between persons and items when there is a vector of hypothetical latent traits (Reckase, 1997, 2009).

The second assumption of IRT is the concept of local independence (LI). LI states that the responses to items are conditionally independent from each other and only depend on the latent trait (de Ayala, 2009; Embretson & Reise, 2000; Hambleton et al., 1991; Lord & Novick, 1968). In other words, the response to one item does not depend, or influence, the response to another after conditioning on the latent trait. Due to LI, item and person characteristics are independent of each other. But, there is a potential for LI to be violated.

The third assumption of IRT models relates to the functional form of the model (de Ayala, 2009). IRT models generally follow an explicit mathematical function used to produce an item characteristic curve (ICC). The mathematical form specifics the number of item parameters to be estimated and used in specifying the ICC. The form of the ICC expresses the direct relationship between the probability of a specific response to precise changes in the latent trait and item's properties (Embretson & Reise, 2000; Hambleton, van der Linden, & Wells, 2010). Thus, item and person characteristics can be placed on the same continuum. Thus, they can be used to predict the probability of a specific response from an individual as well as estimate an individual's ability from their response pattern (Gershon, 2005; Weiss & Vale, 1987).

Unidimensional Dichotomous Models

Often, a single dominant trait underlies examinee performance and the item types require binary scoring (i.e., items are scored as either correct, 1, or incorrect, 0). An example of this

item type is multiple-choice (MC) items. Four primary IRT models exist to represent the relationship between item and person parameters in this case. The most general model is the three-parameter logistic (3PL) model (Birnbaum, 1968). This model includes three item parameters for each item, *i*; these parameters represent item difficulty (β_i), item discrimination (α_i), and a lower asymptote (γ_i; i.e., pseudo-guessing parameter). In the 3PL model, the probability of a correct response on a given item *i* by person *s*, $P(X_{is} = 1)$, is:

$$P_{si}(\theta) = P(X_{si} = 1|\theta_s, \beta_i, \alpha_i, \gamma_i) = \gamma_i + (1 - \gamma_i)\frac{\exp[\alpha_i(\theta_s - \beta_i)]}{1 + \exp[\alpha_i(\theta_s - \beta_i)]}, \quad 2.1$$

where θ_s represents the trait level (i.e., ability) of person *s* (Birnbaum, 1968; Embretson & Reise, 2000). The inclusion of the subscript *i* on item parameter allows for item differences in difficulty, discrimination, and guessing.

In the 3PL model, the item difficulty parameter (β_i) is the point of inflection of the ICC on the ability scale, where the probability of correctly answering the item is $\frac{(1+\gamma_i)}{2}$ (Harris, 1989). Higher β_i values (i.e., more positive) represent harder items, whereas lower values represent easier items. Item discrimination (α_i) represents an item's ability to differentiate between groups of people along the ability continuum, and is the slope of the ICC. Higher α_i values indicate higher discriminatory power; thus, the ICC for the item will be steeper and the item will be more informational when discriminating between various groups around the item's difficulty level (Harris, 1989). Lastly, by including a lower asymptote, the 3PL model provides information for low ability examinees who have a probability greater than 0 of solving specific items (e.g., items that would be considered too difficult for them). The model accounts for the fact that examinees can respond to an item correctly at a level greater than chance without explicitly knowing the correct answer (i.e., guessing; Birnbaum, 1968).

Although the 3PL model allows for unique lower asymptotes, these asymptotes can be constrained to be equal (γ). By eliminating the lower asymptote (i.e., γ = 0) in Equation 2.1, the two-parameter logistic (2PL; Birnbaum, 1968) model can be specified, as follows:

$$P(X_{si} = 1|\theta_s, \beta_i, \alpha_i) = \frac{\exp[\alpha_i(\theta_s - \beta_i)]}{1 + \exp[\alpha_i(\theta_s - \beta_i)]} . \quad 2.2$$

In the 2PL model, items vary on item discrimination and item difficulty. The item difficulty parameter (β_i) is still the point of inflection of the ICC, but this point is now where the probability of correctly answering the item is 50% (Harris, 1989).

It may be plausible that, although items discriminate between groups, each item has the same discriminatory power. This scenario is represented by the one-parameter logistic (1PL) model, in which item discrimination is freely estimated but constrained to be equal across items, as shown below:

$$P(X_{si} = 1|\theta_s, \beta_i, \alpha) = \frac{\exp[\alpha(\theta_s - \beta_i)]}{1 + \exp[\alpha(\theta_s - \beta_i)]} . \quad 2.3$$

Although all items are equally discriminating, they still vary in the location along the ability continuum and can discriminate between individuals using the item's difficulty location (Harris, 1989). Thus, some items may discriminate well among low-ability examinees whereas other items discriminate well among high-ability examinees. The probability of a correct response is still located at 50%, just as in the 2PL model, when $\theta_s - \beta_i = 0$. While equal discriminations may hold for some tests, it is often unlikely this will occur in all settings, especially those related to educational testing (Hambleton et al., 2010).

Lastly, the Rasch model is extremely similar to the 1PL model, but item discrimination is assumed to be one instead of freely estimated. The Rasch model is shown below:

$$P(X_{si} = 1|\theta_s, \beta_i) = \frac{\exp(\theta_s - \beta_i)}{1 + \exp(\theta_s - \beta_i)}, \quad 2.4$$

where, in this model, items only differ in item difficulty. This model is a change in scale from Equation 2.3. The models presented above are expressed in terms of the logistic function and use a logistic (log) metric. Often, the logistic function is utilized over a normal ogive function due to their simplicity and computational advantages (Bock, 1997; Birnbaum, 1968). However, a normal metric is approximated by including a multiplier of 1.7 in the logistic function exponent (i.e., for the 3PL model, $1.7\alpha_i(\theta_s - \beta_i)$; Birnbaum, 1968). Other item types beyond dichotomous items can be used. For example, items with partial credit can be used within polytomous IRT models. However, binary models are only examined.

The amount of information an item *i* contains at specific trait levels, θ, along the continuum can be calculated; this is called the Fischer information of the item (*FI*; Birnbaum, 1968; Embretson & Reise, 2000). *FI* is calculated using the general information equation below for a binary IRT model:

$$I(\theta) = \frac{P'_{si}(\theta)^2}{P_{si}(\theta)Q_{si}(\theta)}, \quad 2.5$$

where $P_{is}(\theta)$ is the conditional probability of answering item *i* correctly, $P'_{is}(\theta)$ represents the first derivative of the conditional probability function at a particular θ, and $Q_{is}(\theta)$ is the probability of an incorrect response ($1 - P_{is}(\theta)$). A test's information (i.e., a set of items) can be calculated by summing the information across all the items, as shown below:

$$TI(\theta) = \sum_{i=1}^{I} I(\theta). \quad 2.6$$

Test information is important for the estimation of latent traits, as will be discussed later.

An item's psychometric properties can affect the amount of information an item provides. Increases in information result when an item's difficulty (β_i) is closer to the examinee's trait level (θ). The more discriminatory power (α_i) an item has also results in higher information values. Lastly, the closer the lower asymptote (γ_i) is to 0, the more information the item provides (Hambleton et al., 1991).

Ability Estimation

IRT attempts to estimate the position of an examinee on the latent trait continuum by using the examinee's behavior on a set of items. The behavior of interest is the examinee's scored responses to the items. For example, for dichotomously scored items, a response pattern can be obtained for each person. Thus, IRT can obtain estimates for the response patterns previously discussed for examinees A and B, who answered the same total number of items correctly but different individual items. based on the specific items endorsed by each examinee. Likelihood estimators (e.g., maximum likelihood estimation, weighted likelihood estimation) and Bayesian estimators (e.g., Expected A Posteriori estimation) can be used to obtain these measurements.

Maximum Likelihood Estimation. Maximum likelihood estimation (MLE; Birnbaum, 1968) is an approach to trait estimation that utilizes the examinee's response patterns to find the value of θ_s that maximizes the likelihood of the pattern. MLE assumes that item psychometric properties are known and item responses and examinee characteristics are independent.

First, the conditional likelihood of a specific response pattern is obtained using Equation 2.7:

$$L(x_{s1}, x_{s2}, \ldots, x_{sI}|\theta_s) = \prod_{i=1}^{I} P_{si}(\theta)^{x_{si}} Q_{si}(\theta)^{1-x_{si}} \qquad 2.7$$

where x_{si} is the observed item response, $P_{si}(\theta)$ is the probability of a correct response, and $Q_{si}(\theta)$ is the probability of an incorrect response (i.e., 1-$P_{si}(\theta)$; Embretson & Reise, 2000). Since these probabilities range from 0 to 1, their product can become an extremely small number. To deal with this issue, the natural logarithm is taken and the log-likelihood function in Equation 2.8 is maximized instead:

$$lnL(x_{s1}, x_{s2}, \dots, x_{sI}|\theta_s) = \sum_{i=1}^{I} x_{si} \ln[P_{si}(\theta)] + (1 - x_{si})\ln[Q_{si}(\theta)] \qquad 2.8$$

The value of θ that maximizes the log-likelihood in Equation 2.8 is the same value that would maximize the likelihood in Equation 2.7 (de Ayala, 2009; Embretson & Reise, 2000).

This calculation can be quite cumbersome and Newton-Raphson is often employed to find the value of θ. To use this procedure, the first $\left(\frac{\partial lnL}{\partial \theta}\right)$ and second $\left(\frac{\partial^2 lnL}{\partial \theta^2}\right)$ derivatives of the log-likelihood function, based on the IRT model being used, must be calculated. Estimates of ability, $\hat{\theta}$, are updated using an iterative process and these derivatives. An initial estimate ($\widehat{\theta_0}$) is first given to an examinee as an approximation of their true latent trait level; this estimate can be determined using prior information about the examinee or can be equal for all examinees. Using t$\widehat{\theta_0}$, the first and second derivatives are calculated and a ratio (ε) of the derivatives is obtained ($\varepsilon = \left.\frac{\partial lnL}{\partial \theta}\right/\frac{\partial^2 lnL}{\partial \theta^2}$). An updated estimate is obtained by subtracting this ratio, ε, from the previous estimate ($\widehat{\theta_1} = \widehat{\theta_0} - \varepsilon$). The standard error of measurement for a specific MLE θ estimate is calculated as follows (de Ayala, 2009; Embretson & Reise, 2000):

$$SE(\hat{\theta}) = \frac{1}{\sqrt{TI(\theta)}} \qquad 2.9$$

MLE has some positive features. It is an unbiased estimate of θ and is an asymptotically efficient estimator with normally distributed errors (Birnbaum, 1968; Embretson & Reise, 2000).

However, MLE may also pose problems to trait estimation. It assumes the item responses fit the model. Local maxima, instead of global maxima, may be achieved in certain situations. However, a major problem with MLE is that the algorithm cannot provide a trait estimate for response patterns where all items are answered correctly or all items are answered incorrectly (i.e., perfect response patterns). These patterns produce monotonically increasing or decreasing likelihood functions, respectively; these functions will have no absolute maximum. The ability estimates will be either $\theta_i = +\infty$ or $\theta_i = -\infty$. Thus, in this instance, another approach may be used or boundaries may be placed on the estimates.

Weighted Likelihood Estimation. Lord (1983) showed that MLE is biased when estimating ability, especially at the extremes of the latent continuum, and proposed a way to remove the first order bias term. In order to correct for this bias, Warm (1989) added a weighting factor to correct for the noted bias and called it a weighted likelihood estimate (WLE). In his approach, he utilizes the first derivation of the MLE likelihood function:

$$\frac{\partial lnL}{\partial \theta} = \frac{\sum_{i=1}^{I}(x_{si} - P_{is}(\theta))P'_{si}(\theta)}{P_{si}(\theta)Q_{si}(\theta)} = 0\,, \qquad 2.10$$

where all terms have been previously defined. In order to obtain an unbiased estimate, an estimate of θ must satisfy Equation 2.11 (Warm, 1989):

$$\frac{\partial lnL}{\partial \theta} = \frac{\sum_{i=1}^{I}\left(x_{si} - P_{is}(\theta)\right)P'_{si}(\theta)}{P_{si}(\theta)Q_{si}(\theta)} + \frac{J}{2I} = 0$$
$$where\ J = \frac{\sum P'_{si}(\theta)P''_{si}(\theta)}{P_{si}(\theta)Q_{si}(\theta)}\ and\ I = \frac{P'_{si}(\theta)^2}{P_{si}(\theta)Q_{si}(\theta)}\,. \qquad 2.11$$

WLE is attractive for many tests, as it produces an estimate that is less biased than that produced by MLE.

Maximum A Posteriori Estimation. As stated previously, MLE estimation cannot provide latent trait estimates when the examinee answers all items either correctly or incorrectly; some variation in the responses are necessary. One way to combat this limitation is to introduce a prior distribution to the estimation process. As long as the test administrator/researcher is comfortable assuming the estimated value falls within a specified range, the prior gives more efficient information and will protect against the influence of outliers (Embretson & Reise, 2000). Maximum A Posteriori (MAP) estimation, also known as Bayes Model Estimation, is a Bayesian estimation procedure that places a prior distribution on the person estimates. The goal of MAP is to determine the value of θ that maximizes the posterior distribution, or the mode (Bock & Aitkin, 1981).

Similar to MLE, MAP is an iterative procedure and follows most of the same steps. Thus, it is necessary to have an initial estimate of the examinee's true latent trait level ($\widehat{\theta_0}$). Using the examinee's response pattern and the item's psychometric properties, the log-likelihood is calculated. Also, as with MLE, the first and second derivatives of this log-likelihood at the initial trait estimate need to be computed. However, before finding ε, the derivatives are adjusted by incorporating the prior distribution. This prior distribution is a hypothetical distribution from which the examinees are a random sample. A common prior distribution is the standard normal distribution, $\theta \sim N(0,1)$, but any *appropriate* distribution can be used. It is important to emphasize the use of an appropriate distribution. If the prior distribution is inappropriate, the resulting trait estimates could be biased and misleading (Embretson & Reise, 2000). A posterior distribution results from the multiplication of the log-likelihood by the prior distribution. At this point, ε is calculated and an updated trait estimate is obtained. Standard errors for MAP are calculated similar to MLE, but information is from the posterior distribution.

Inclusion of a prior distribution into the estimation of θ increases precision of the estimate, since more information is utilized. It also allows for θ estimation of all examinees, regardless of their response pattern. However, the estimates can be biased, as the expected value of the MAP estimate does not equal the true value. The addition of the prior distribution results in estimates that are often pulled towards the mean of the prior distribution. Shorter tests are more influenced by the presence of a prior distribution. Lastly, as previously stated, if an inappropriate prior distribution is chosen, the resulting estimates may be even more biased (Embretson & Reise, 2000).

Expected A Posteriori Estimation. Expected A Posteriori (EAP) estimation finds the mean of the posterior distribution (Bock & Aitkin, 1981; Embretson & Reise, 2000). It is also known as the Bayes Mean Estimate (de Ayala, 2009). EAP uses a discrete set of probability weights for a fixed set of θ values. An EAP estimate is obtained using Equation 2.12, below:

$$\theta_s = \frac{\sum_{r=1}^{R}[Q_r \times L(Q_r) \times W(Q_r)]}{\sum_{r=1}^{R}[L(Q_r) \times W(Q_r)]}, \qquad 2.12$$

where *Qr* represents the quadrature nodes for the fixed set of θ values chosen, *W(Qr)* represent the discrete weights for each quadrature node, and *L(Qr)* is the exponent of the log-likelihood function at each of the *r* ($r = 1, \ldots, R$) quadrature nodes. Generally, the nodes and weights represent a standard normal prior distribution, $\theta \sim N(0,1)$, but as with MAP, other appropriate distributions can be chosen. Standard error for the EAP estimate is calculated using Equation 2.13.

$$SE = \sqrt{\frac{\sum_{r=1}^{R}[(Q_r - \theta)^2 \times L(Q_r) \times W(Q_r)]}{\sum_{r=1}^{R}[L(Q_r) \times W(Q_r)]}}. \qquad 2.13$$

EAP also yields estimates for all possible response patterns. It is a non-iterative approach and is computationally faster, making it advantageous over other Bayesian estimators that use the mode of the posterior distribution. It is also easy to use with both dichotomous and polytomous models. However, EAP estimates may be biased and regressed to the mean. This is overly apparent when an inappropriate prior is chosen.

Examinee Classification

Two main goals of testing are to determine an examinee's ability with minimal error and to assign an individual examinee to a category that represents the level of skill proficiency as measured by the test (Birnbaum, 1968; Hambleton et al., 1991). This score can be used in conjunction with a reference group to provide meaning to the examinee's location (e.g., scores in the upper 20%). However, this score can also be used to place the examinee into various groups, or categories, using cutscores. This process is called classification and can be used to categorize individuals into two or more groups (e.g., master versus non-master; basic, proficient, or advanced). IRT allows for classification based on θ estimates of examinees using the full response pattern of the individual.

Mastery testing is a type of testing that uses IRT-based θ to classify examinees into various groups (Lord, 1980). To classify individuals into groups, cutscores (θ_C) are computed based on true-score levels that define mastery. A test may have only one cutscore (θ_C) that defines the difference between two groups, such as masters and non-masters. If an examinee's $\hat{\theta}$ is below θ_C, the examinee is classified as a non-master; if it is above θ_C, the examinee is classified as a master. The same approach can be used for multiple cutscores (θ_{C1} and θ_{C2}) where examinees are classified into multiple groups (e.g., basic, proficient, advanced). Classification into the basic and advanced groups would follow a similar pattern to the two group classification.

The difference would be in classifying an examinee into the proficient category. To be classified as proficient, an examinee's $\hat{\theta}$ would need to be greater than θ_{C1} but equal to or less than θ_{C2} (Lord, 1980).

IRT allows for the estimation of two separate classification indices. The first, classification consistency, is the probability that an examinee with a specified θ would be classified into the same category on separate administrations of an assessment (Lee, 2010). Classification accuracy is the rate at which examinees are classified into their true category based on their "true" ability (Lathrop & Cheng, 2013; Lee, 2010). False positives occur when an examinee is classified into a higher category than their true category, and false negatives occur when an examinee is classified into a lower category than their true category (Lee, 2010; Stone, Weissman, & Lane, 2005).

Classification via IRT requires multiple considerations. First, the choice of IRT model will affect the classification of examinees. Stone et al. (2005) found that more consistent classifications were obtained when the IRT model fit the data. In this study, a 3PL model resulted in more consistent examinee classification than a 1PL model did when using multiple-choice items. Multiple-choice items often result in higher levels of guessing; the 3PL model includes a parameter for this item characteristic and may be more appropriate when representing items. Thus, they found that the 1PL model systematically underestimated ability estimates, which affected classification.

Another consideration is the location of the cutscores. Classification accuracy is conditional on the placement of the cutscore (Lathrop & Cheng, 2013). Lathrop and Cheng (2013) found that classification accuracy was low when the cutscore was at the mean of the ability distribution, but increases as the cutscore moved further away from the mean. Lower

classification accuracy is also obtained at locations along the ability distribution where the standard error of measurement is high. This is because less information is available at these locations and results in more classification errors (Lathrop & Cheng, 2013).

Lastly, a choice of what to base the classification decision on affects accuracy. Classification can be based on a total score, *x*, for the test or it can be based off the latent trait estimate, $\hat{\theta}$, obtained via IRT. If using the Rasch model, the total score is a sufficient statistic and results in high classification accuracy. However, if a model is chosen where total score is not a sufficient statistic, such as in the 3PL model, $\hat{\theta}$ is preferable (Lathrop & Cheng, 2013). This decision relates to the first consideration of choosing the appropriate model to represent the data.

Test Designs

Tests are generally constructed with a specific purpose (Crocker & Algina, 2008; Guilford, 1954). Determining the purpose includes establishing the construct of interest (e.g., mathematics proficiency, aptitude, personality), the population of interest (e.g., high school students, job applicants, military personnel), and the behaviors that are representative of the construct (e.g., solving mathematics items, solving pattern sequences, responding to agree/disagree items). Other features relating to the purpose of the test is whether the test will discriminate among a broad or narrow range of abilities based on the goal of the test (e.g., examinee rank or examinee classification). The last decision is how the final score will be used (i.e., what gives the final score meaning). One approach is to interpret an individual examinee's score against a representative group; this is considered norm-referenced measurement. Another approach is to gain an absolute level of performance for the examinee; this is criterion referenced measurement.

A final determination for a test is whether it is to be administered adaptively or not. Non-adaptive tests, or conventional tests (CTs), are those in which a fixed set of items is administered to all examinees irrespective of administration medium, such as paper or via computer (Garrison & Baumgarten, 1986; Weiss, 1985); conventional tests may also be called linear fixed length test (LFTs). These tests may be scored using traditional classical test theory (CTT) statistics, where an observed number correct score is obtained for each examinee, or using IRT methods. However, a major issue exists with the use of conventional tests; this issue is ability-difficulty mismatches (Garrison & Baumgarten, 1986; Mead & Drasgow, 1993). While this issue is easy to understand, it is a serious disadvantage. Ability-difficulty mismatches occur when items on the assessment are either too easy or too hard for the examinee (Mead & Drasgow, 1993). Thus, a conventional test may not accurately reflect the true ability of an individual if items do not exist around his or her ability. Consequently, the standard error of measurement for an examinee with a large ability-difficulty mismatch might be quite high (Garrison & Baumgarten, 1986). A solution to this issue is the use of adaptive (e.g., computer adaptive or multistage) testing designs.

Adaptive testing is a "process of test administration in which test items are selected for administration on the basis of the examinee's responses to previously administered items" (Weiss & Kingsbury, 1984, p. 361). This approach to measurement resolves the ability-difficulty mismatch issue plaguing conventional tests by tailoring each assessment to the examinee (Weiss, 1985). Thus, examinees only receive items appropriate for their ability level, creating an ability-difficulty match (Chang, 2014). Two approaches to adaptive testing are computerized adaptive tests (CAT) and multistage tests (MST).

CAT

Computerized adaptive testing (CAT) is one type of adaptive test design that tailors an assessment to the individual, resulting in a test that is optimal for an examinee with a specific ability level, θ. These assessments utilize IRT, since item and people characteristics are placed on the same latent continuum. CATs can be designed to measure achievement, aptitude, or personality traits. During the CAT process (Figure 2.1), an initial item from the item pool is administered to an examinee, and based off their response, a provisional estimate of ability is obtained. Using this estimate, $\widehat{\theta_1}$, the next best item is chosen from the pool of items and administered to the examinee. This item is chosen to provide the most information conditional on the examinee's current ability estimate. Based on the response to this new item and the previous one, an updated estimate, $\widehat{\theta_2}$, is obtained. The process continues in this "item, response, update" fashion until the end of the test and a final estimate is obtained using information from all items and item responses (Weiss & Kingsbury, 1984). These tests are structured such that final ability estimates are obtained such that all examinees achieve a similar percentage correct score (e.g., examinees answer approximately 50% of the items administered correctly; Bergstrom, Lunz, & Gershon, 1992). The goal of a CAT can be estimation, in which a precise estimate of proficiency in the domain is desired, or classification, in which the goal is to make a decision regarding the categorization of an individual (Eggen, 2011).

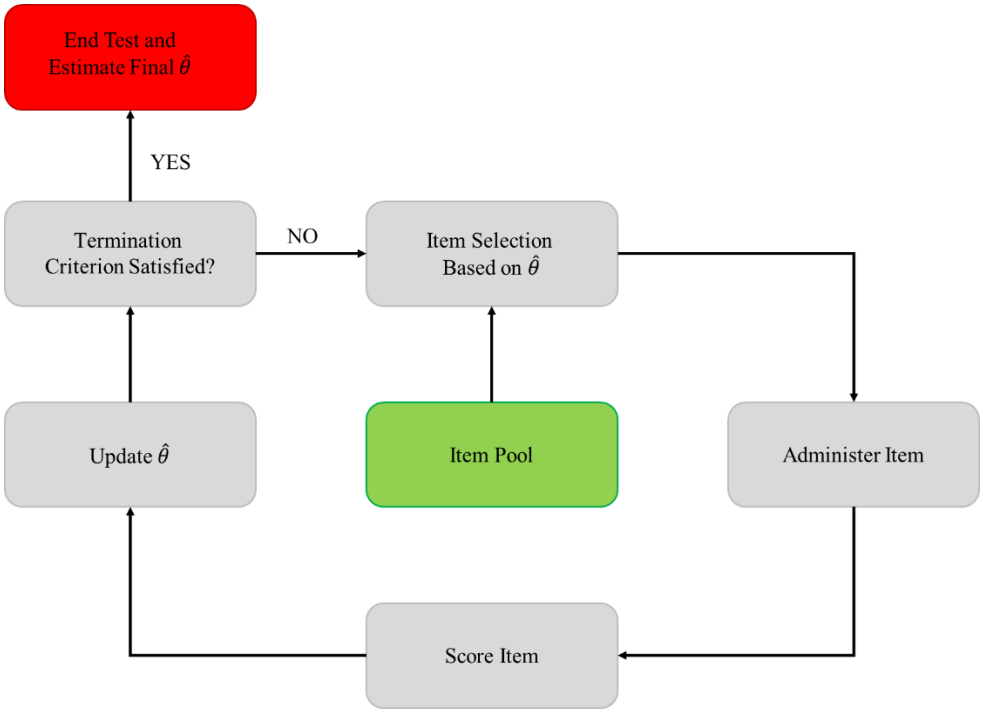

Figure 2. 1. Example of a CAT.

A CAT may be terminated when a pre-specified level of measurement precision has been attained, after a fixed number of items has been administered, or after a specific time interval has elapsed (Thissen & Mislevy, 2010). CATs are designed to administer the shortest test possible to examinees. Thus, using equiprecise measurement, the standard error (SE) of $\hat{\theta}$ is examined and when a preset level is obtained, testing stops. This level can vary for each test and depends on the test's purpose (Weiss & Kingsbury, 1984; Weiss, 1982; Weiss & Vale, 1987). This approach to termination results in variable-length CATs (i.e., test length can vary between examinees). Another approach to termination is a fixed-length CAT, in which a specific number of items is administered to all examinees and the assessment stops when the last item has been reached (Gershon, 2005). The last approach stops the test after a fixed amount of time (e.g., 2 hours) has elapsed (Thissen & Mislevy, 2010). Variable-length tests usually result in less measurement error and are more efficient than fixed-length tests (Babcock & Weiss, 2012). In

practice, a combination of these approaches is usually considered and implemented for practical, political, and legal reasons (Babcock & Weiss, 2010; Gershon, 2010). For example, examinees may complain they received a longer (or shorter) test than another examinee and suggest the test is unfair. Or, if using equiprecise measurement, an examinee may run the risk of having every item administered to them during the testing window. Thus, it is pertinent to think of possible issues that might arise and use the best combination of termination rules (Gershon, 2005). Babcock and Weiss (2012) recommend using a combination of variable-length termination criteria (e.g., low standard error) as well as a minimum number of items administered constraint.

Since a CAT tailors the assessment to the examinee by selecting the next best item for the examinee's provisional estimate, an item selection method must be chosen. An item selection method specifies how the CAT chooses items for the examinee (Meijer & Nering, 1999). Random selection of items would result in longer, less precise assessments and would introduce an ability-difficulty mismatch issue as in conventional tests. One approach (i.e., point information criterion) to item selection is to choose the most informative item at the examinee's current $\hat{\theta}$. Using this provisional estimate, Fischer's information (i.e., Equation 2.5) would be computed and the next item chosen such that it maximizes this information at the estimate (Gershon, 2005; Lord, 1980; Thissen & Mislevy, 2010; Thompson & Weiss, 2011; Weiss, 1982; Weiss & Kingsbury, 1984). More informative items reduce the error of measurements at $\hat{\theta}$ (Weiss & Vale, 1987). However, issues do exist with this approach, especially when there is no variation in the examinee's response pattern. Veerkamp and Berger (1997) proposed the likelihood weighted information criterion, which uses the likelihood function as a weight in item selection when it is more likely the item's information function is close to the examinee's true ability. In this approach, ability is only estimated once at the end of the test administration.

However, it solves the issue with the point information criterion. Another approach would be to select the item with the smallest posterior variance of $\hat{\theta}$; this approach is appropriate if a Bayesian method (e.g., MAP, EAP) is used (Meijer & Nering, 1999; Weiss & Kingsbury, 1984). Both of these approaches are similar and would result in a similar set of administered items.

Ability estimation. When estimating ability, CTT can be used within a CAT framework, but IRT approaches provide more precision and are often preferred (Thompson & Weiss, 2011). As previously discussed, MLE and WLE can only be used when mixed response patterns, in which correct and incorrect responses have been recorded, are obtained, which is a drawback to these approaches. Multiple studies have been conducted comparing likelihood based estimators and Bayesian estimators in terms of various error indices (i.e., bias, standard error (SE), and root mean square error (RMSE)) in multiple situations. These estimators have been investigated in relation to various termination rules, such as fixed length and variable length termination (Doebler, 2012; Gorin, Dodd, Fitzpatrick, & Shieh, 2005; Wang, Hanson, & Lau, 1999; Wang & Wang, 2001; Yi, Wang, & Ban, 2001). They have also been investigated with dichotomous (Doebler, 2012; Wang et al., 1999; Yi et al., 2001) and polytomous (Gorin et al., 2005; Wang & Wang, 2001) IRT models. Item pools have also been varied in order to examine the impact of item bank size (small versus large) and item bank distribution (peaked versus rectangular) on ability estimators (Doebler, 2012; Gorin et al., 2005; Wang et al., 1999; Wang & Wang, 2001; Yi et al., 2001).

Studies often compare the four ability estimators listed above. However, Wang et al. (1999) introduced essentially unbiased Bayesian estimators. These essentially unbiased Bayesian estimators (EU-EAP and EU-MAP) use a beta prior specifically designed to reduce the bias commonly seen in these estimators without increasing the SE. Thus, in contrast to the

normally used normal prior that includes information reflecting the examinee population, the essentially unbiased prior solely serves to decrease bias and does not reflect any prior information about the examinee. Gorin et al. (2005) examined EAP estimators with various priors, resulting in a total of six ability estimators: MLE, WLE, EAP with a uniform prior (EAP-U), EAP with a normal prior (EAP-N), EAP with a negatively skewed prior (EAP-NS), and EAP with a positively skewed prior (EAP-PS). EAP-U was considered an uniformed prior, as no prior information was used, whereas the other EAP estimates had informed priors (i.e., EAP-N, EAP-NS, EAP-PS). A variable termination rule was used. Informed versus uniformed priors were examined for their effects on error indices.

For tests using a fixed-length termination rule and dichotomous IRT models, bias can be seen across all estimators if the test is long enough. Bayesian estimators often produce bias towards the mean of the prior (Meijer & Nering, 1999; Weiss & Kingsbury, 1984). Likelihood estimators produce a different pattern: ability is often overestimated for high performing examinees and underestimated for low performing examinees (Meijer & Nering, 1999). In terms of specific ability estimators, WLE has less bias than MLE and EAP has less bias than MAP (Doebler, 2012). WLE and EU-MAP produce similar results; these estimators were less biased than MAP, and their bias was lower or equal to bias in MLE and EU-EAP (Wang et al., 1999). In terms of SE, the estimators are listed in order of increasing error: MAP, EU-EAP, EU-MAP, and MLE. WLE estimators were similar to EU-MAP estimates in terms of SE except at the extreme low end of the ability continuum, in which WLE produced slightly higher SEs. For RMSE, EU-EAP was lower than EU-MAP, which was lower than MLE. Again, WLE performed similarly to EU-MAP except at the extreme low end. Thus, WLE, EU-EAP, and EU-MAP perform better in terms of bias and SE than MLE (Wang et al., 1999). Similar results for

these estimators are obtained for polytomous IRT models and a fixed-length termination rule (Wang & Wang, 2001). WLE produced the smallest bias over the largest ability range. MLE produces outward bias (e.g., overestimation and underestimation at extremes) while EAP/MAP produce inward bias (e.g., bias toward mean). WLE has the smallest SE in the middle ability ranges, but MLE is close. EAP and MAP have lower SEs than MLE/WLE at the ability extremes (Wang & Wang, 2001).

Comparison of the ability estimators with variable-length termination rules produce different results. Under a variable-length termination rule for dichotomous IRT models, the likelihood estimators and MAP estimator were strongly biased; WLE results in greater bias than MLE (Wang et al., 1999; Yi et al., 2001). EU-EAP, EU-MAP, and EAP resulted in the lowest bias across the estimators. Similar patterns were found for SE and RMSE. The likelihood estimators resulted in larger SEs and RMSEs than the Bayesian estimators, with the exception of the extremes, in which MAP had the largest RMSEs (Wang et al., 1999; Yi et al., 2001). Using a polytomous IRT model, EAP estimates with informed priors showed bias towards the mean, as expected, and likelihood estimators also produced inward bias (Gorin et al., 2005; Wang & Wang, 2001). WLE was not an improvement over MLE in terms of bias (Wang & Wang, 2001). All Bayesian estimators resulted in lower SEs than WLE and MLE (Gorin et al., 2005; Wang & Wang, 2001). EAP-U had higher SEs than EAP estimators with informed priors. Thus, in general, ability estimates were more stable for EAP estimators with informed priors over uniformed priors. MLE and WLE were comparable in terms of standard error, so WLE was not an improvement over MLE (Gorin et al., 2005).

When taking these results as a whole, a few general conclusions can be drawn. The choice of the CAT termination rule greatly influences which ability estimator should be chosen.

For fixed length tests, WLE is an improvement over MLE and Bayesian estimators in terms of bias (i.e., it reduces bias more). However, it may be non-convergent when response patterns do not include mixed responses, as described earlier. In this case, Bayesian estimates may be utilized for the whole test or until a mixed pattern is observed (Gorin et al., 2005). For variable length tests, WLE is not an improvement over MLE and may increase bias. Informed Bayesian estimates improve estimation, but uninformed Bayesian estimates are not extremely detrimental. Thus, there is a trade off when choosing an ability estimator for an assessment. The likelihood estimators (MLE and WLE) are often less biased but the Bayesian estimators (EAP and MAP) are often more precise. If bias is a concern, likelihood estimators or the essentially unbiased Bayesian estimators are preferred; if SE is a concern, Bayesian estimators are preferred (Wang et al., 1999).

Classification. Computerized classification tests (CCTs) are a special form of CAT in which the goal of the test is to adaptively administer a test such that an examinee can be classified into mutually exclusive categories based on the relationship of the ability estimate to a cutscore. CCTs maximize the efficiency of the test by having small classification errors and reduced items (Eggen, 2011; Gnambs & Batinic, 2011; Nydick, 2014; Thompson, 2007, 2009; Weiss, 1982). Thompson (2007) recommends using variable-length computerized classification tests (VL-CCT) as the name for CCTs that terminate after a classification decision can be made, thus resulting in variable-length tests. CCTs may also be designed to administer a fixed number of items. However, for this discussion, CCT will be retained.

One difference between standard CAT designs and CCT is the termination criterion utilized. While CCTs can be terminated after a fixed number of items are administered, this approach is often not optimal. Therefore, several approaches exist for termination. One

approach is an ability confidence interval approach (ACI; Patton, Cheng, Yuan, & Diao, 2013; Thompson, 2009; Weiss & Vale, 1987). In this approach, an estimate of ability for examinee *s*, $\widehat{\theta_s}$, is obtained and a confidence interval (CI) is constructed around the estimate. The confidence interval is calculated using Equation 2.14.

$$I = \widehat{\theta_s} \pm z_{\frac{1-\alpha}{2}} * SE\,. \qquad 2.14$$

In this equation, $z_{\frac{1-\alpha}{2}}$ corresponds to the (1-α) CI (Patton et al., 2013; Thompson, 2007). An examinee is classified into a category when the cutscore, $\hat{\theta}_C$, is not contained within the CI. Tests of varying length are obtained using ACI; if $\widehat{\theta_s}$ falls near the $\hat{\theta}_C$, the examinee will receive a longer test than if a larger discrepancy existed (Patton et al., 2013). For example, if examinees are being classified as either masters or nonmasters, an examinee is classified as a master when their $\hat{\theta}$ and corresponding CI is completely above the cutscore, $\hat{\theta}_C$. When ACI utilizes estimate-based (EB) selection, less items were required to make a classification decision with similar accuracy to cutscore-based (CB) selection (Thompson, 2011). CCTs that used this approach were originally coined *adaptive mastery tests* (Thompson, 2007; Weiss & Kingsbury, 1984), but this title is too restrictive. This approach was usually estimate-based and made dichotomous classifications. ACI can be used in broader applications, such as adaptive testing for assigning grades (Weiss & Kingsbury, 1984).

A second termination approach is similar to that utilized in CAT. This approach uses the SE to end the test, aiming for equiprecise measurement across all examinees (Thissen & Mislevy, 2000). Thus, the test ends when a pre-specified level of measurement has been achieved, and the examinee is classified into a category. This conditional standard error (CSE) rule might cause tests to end prematurely. While a pre-specified level of error has been

achieved, classification accuracy may be impacted because ability estimates might be biased around extreme cutscores (Patton et al., 2013).

A third, and more common, termination approach in CCT is the sequential probability ratio test (SPRT). The SPRT phrases the problem in terms of a ratio of the likelihoods of two hypotheses. These hypotheses are shown below in terms of a dichotomous classification (Eggen, 2011; Nydick, 2014):

$$\begin{aligned} H_0&: \theta_1 = \theta_C - \delta \\ H_1&: \theta_2 = \theta_C + \delta\,, \end{aligned} \tag{2.15}$$

where δ represents equally-spaced indifference regions around the cutscore θ_C. These indifference zones reflect the idea that making accurate decisions for individuals close to the cutscore cannot be guaranteed due to measurement error (Eggen, 2011). Larger indifference regions may decrease the number of items administered on an assessment, but classification accuracy may suffer (Thompson, 2011). A likelihood ratio test is computed (Eggen, 2011; Nydick, 2014; Thompson, 2007, 2011):

$$LR(\theta_2;\theta_1) = \frac{\prod_{i=1}^{I} P_{2i}(\theta)^{x_{si}} Q_{2i}(\theta)^{1-x_{si}}}{\prod_{i=1}^{I} P_{1i}(\theta)^{x_{si}} Q_{1i}(\theta)^{1-x_{si}}}\,. \tag{2.16}$$

This ratio is then compared to two decision points with acceptable error rates, where α represents Type I error rate and β represents Type II error rate, shown below (Thompson, 2007):

$$\begin{aligned} Lower\ decision = \ B &= \frac{\beta}{1-\alpha} \\ Upper\ decision = A &= \frac{1-\beta}{\alpha}. \end{aligned} \tag{2.17}$$

If the ratio is above *B*, the examinee is classified as above the cutscore (reject H_0). If the ratio is below *A*, the examinee is classified as below the cutscore (accept H_0). If the ratio is between *A* and *B*, another item is administered and the process is continued (Eggen, 2011; Thompson, 2007,

2011). The SPRT can be extended to more than two categories. SPRT can continue infinitely for examinees near the cutscore; thus, it can be adapted to end after a specified number of items, and is known as the Truncated SPRT (TSPRT; Eggen, 2011). Once the maximum number of items has been administered, the most probable decision is made. These termination approaches generally result in shorter tests than ACI (Thompson, 2011). However, within SPRT, EB selection increases the number of items needed to make a classification decision with reduced accuracy (Thompson, 2009).

MST

Multistage testing (MST) is another approach to adaptive testing and can be thought of as a special case of CATs. Where CATs adapt to the individual at an item level, MSTs adapt at the item-set level (Hendrickson, 2007; Yan et al., 2014). In fact, a CAT can be obtained from a MST if each item-set was only composed of one item. A conventional test can be considered a special case of a MST, in which there is only one item-set. MSTs, often called computer adaptive sequential tests (CAST; Luecht & Nungester, 1998; Yan et al., 2014), are also an improvement over conventional tests in relation to the ability-difficulty mismatch (Weiss, 1985). Some researchers think MSTs are "the ideal compromise between linear (nonadaptive) tests and computerized adaptive tests (CATs) in that they allow some of the content and quality controls of linear tests, while providing some of the greater efficiency and flexibility of CATs" (Zwick & Bridgemen, 2014, p. 271). Examinees receive items that are appropriate for their ability level, but they are administered in groups. MSTs offer improved measurement precision and shorter tests when compared to conventional tests. Various versions of multistage tests have been used, such as Cronbach and Gleser's (1965) two-stage sequential test design for selection or rejection of borderline candidates for employment (i.e., classification) and Lord's (1971) two-stage testing

for measurement. MSTs can be PBTs with separate administrations, or they can be CBTs with or without separate administration times (Yan et al., 2014). MSTs can also be adaptive or non-adaptive. However, research on MSTs was eclipsed by the development of CATs.

Like CATs, MST designs are administered via an algorithmic approach. Prior to administration, modules, or groups of items, are assembled; these modules may also be called testlets. Modules can be composed of discrete items, performance exercises, problem-based item sets, common-stem items, or other variations (Leucht, 2014). These modules are selected by the algorithm and administered to examinees. Figure 2.2 presents the basic concept of a MST with three stages; this can be designated as a 1-3-3 MST design. The stages and modules administered together represent a panel, or a complete test. Panels often will not provide the level of precision that a CAT can because the level of adaptability is lower (Chuah, Drasgow, & Luecht, 2006). For each panel, the examinee is administered a routing test in the first stage. Based on the responses of the examinee to the items contained in the routing test, the examinee is routed to either module A, B, or C in stage 2. After answering these items, the examinee is routed to one of the three modules in stage 3; this final module is considered the measurement test. These later stage modules are designed to differentiate between narrower proficiency levels than the routing test (Hendrickson, 2007). The sequence of modules taken through each stage is considered the path. Each of the modules in each stage vary in total difficulty as it relates to the examinee's proficiency, and specific rules govern the path an examinee can take. For example, if the examinee is routed to the easy module A based on his or her responses to the routing test, the examinee can only be routed to either module D or E in stage 3. Even if the examinee answers all items in module A correctly, there is no path to the harder module (i.e., Module F) in stage 3 (Zenisky & Hambleton, 2014; Yan et al., 2014). While this explanation is based on the

three-stage, seven module example MST presented in Figure 2.2, it can be generalized to an *m*-stage, *k*-module MST.

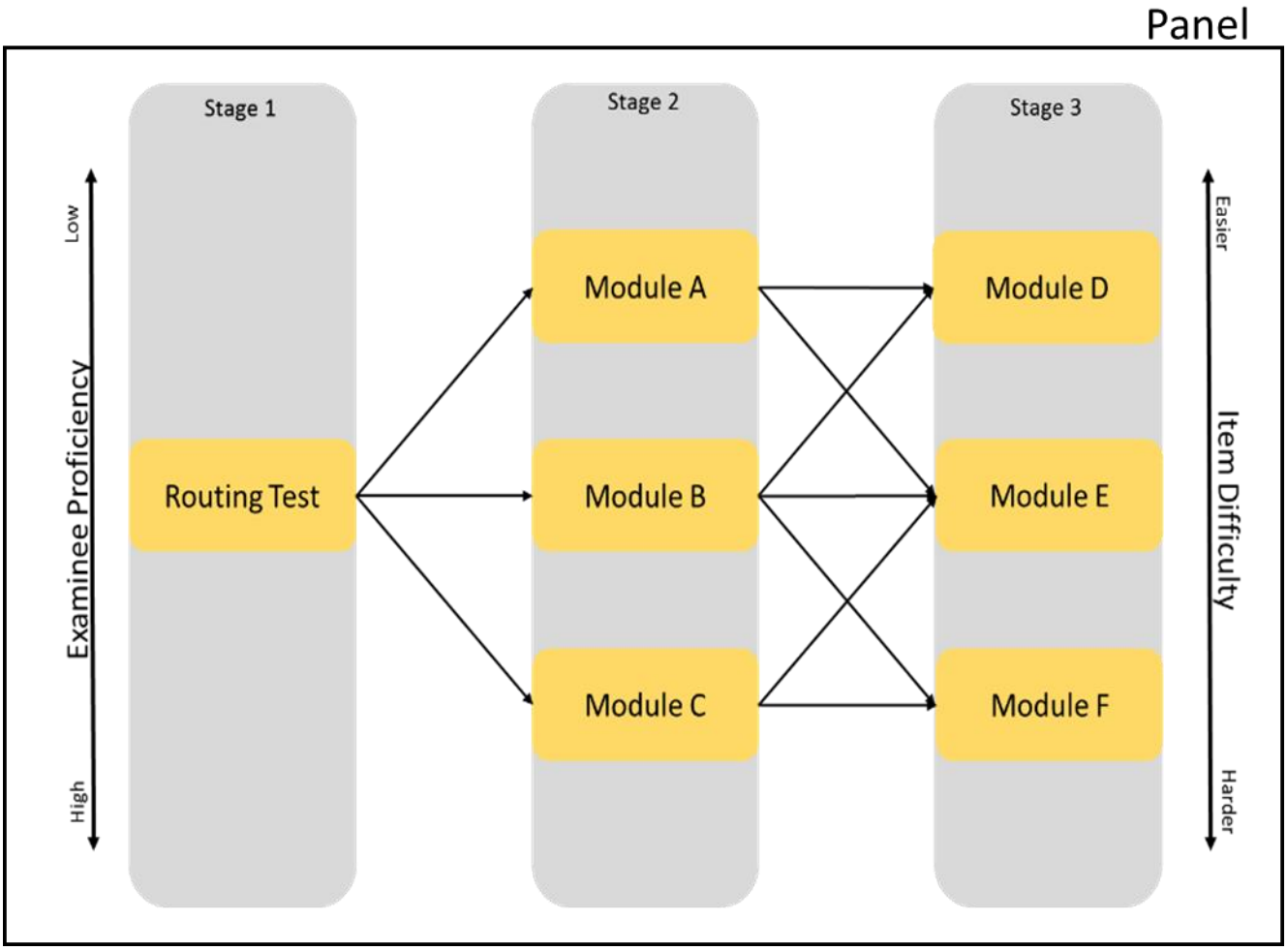

Figure 2. 2. Example of a MST.

As with conventional tests and CATs, the purpose of the MST must be determined prior to development of the modules, stages, panels, and test. The test might be used for criterion referenced measurement or norm-referenced measurement (Hendrickson, 2007; Zenisky & Hambleton, 2014). This decision will affect the routing method chosen to navigate examinees from one module to the next. Since MSTs are composed of various modules at each stage, the module must be located along the difficulty continuum. Often, the routing test in stage 1 is of

average difficulty; this allows for a first pass at estimating an examinee's location (i.e., proficiency) on the latent scale and aids in selecting the next-best module for the examinee (Lord, 1980; Yan et al., 2014). The modules in the successive stages vary in their difficulty. Terminating a MST is different from CATs. Where CATs can be fixed-length or variable-length, MSTs are terminated after a set number of stages have been administered (i.e., the MST has a fixed number of items and stages). Routing in a MST can be done by using a pre-specified proportional schema, in which a specific proportion of examinees are routed to each module; this approach is termed defined population intervals (Luecht et al., 2006; Zenisky & Hambleton, 2014). Information from prior examinees is used to create decision points, θ_ds, such that a specific proportion (i.e., 33%) of the population is exposed to the module. This approach may manage module exposure but may lead to inaccurate estimation if the prior information used to set the θ_ds is inaccurate (Zenisky & Hambleton, 2014). Another routing approach, similar to that used in a CAT, selects the next informative module based on $\hat{\theta}$ (Luecht & Nungester, 1998; Zenisky & Hambleton, 2014).

Ability estimation. In MST, previously discussed trait estimation approaches (e.g., MLE, WLE, MAP, and EAP) can be utilized at the end of the assessment to deliver a point estimate and corresponding SE when item parameters are known. These estimates can also be obtained at the end of each stage to route examinees to the next module (Hendrickson, 2007; Weissman, 2014). Scoring, as well as classification based on the proficiency score, is conducted after an examinee has completed the entire MST (i.e., completed a path through a panel; Weissman, 2014).

Classification. Similar to CCTs, classification in MSTs, called CMSTs, aims to categorize examinees into multiple groups with minimal error (Smith & Lewis, 2014). When

classification is the goal in MST, several decisions must be made. First, the location of the cutscore(s), θ_C, must be determined. If cutscores are located near the densest part of the examinee population (i.e., the mean), classification accuracy might suffer. Thus, classification is more accurate when cutscores are in the extremes of the distribution (Smith & Lewis, 2014). CMSTs can utilize the techniques discussed with CATs (i.e., SPRT, ACI, equiprecise measurement; Smith & Lewis, 2014; Weissman, 2014) to make classification decisions.

Research is convoluted on whether MST designs provide any increase in classification accuracy over conventional tests or CATs. Xing and Hambleton (2004) examined the impact of item pool size and item quality across three testing designs: linear tests, CATs, and MSTs. Results suggest that classification accuracy is higher when the item pool is large and items are more informative, which is a standard result. Interestingly, their results suggest that a MST design functions similarly to a linear test around the cutscore, but does have increased measurement across the whole proficiency continuum. Hambleton and Xing (2006) examined the potential of these test designs when the candidates were centered around the cutscore or not. Results support the use of MST designs when the cutscore and population mean are matched (i.e., optimally designed). They concluded that the choice of test design did not have much influence over classification, although the CAT design did perform slightly better than the other two designs and MSTs perform better than conventional tests (Xing & Hambleton, 2004; Hambleton & Xing, 2006).

Advantages and Disadvantages of Test Designs

Conventional tests have a major limitation in that there is an ability-difficulty mismatch. Both CAT and MST designs remove this limitation by adapting to the examinee (Gershon, 2005; Weiss, 1982, 1985; Weiss & Vale, 1987). CATs and MSTs have many advantages over

conventional tests, both for the test-taker and the test-developer. Advantages for the examinee include easier test administration, greater test availability, and immediate score reporting with the marriage of the assessment with computers (Hendrickson, 2007; Meijer & Nering, 1999; Patsula, 1999). These assessments are often shorter and more efficient at estimating ability, since the exam is tailored to the specific examinee, so comparable scores can be provided in less time with fewer items (Hendrickson, 2007; Meijer & Nering, 1999; Weiss, 1985). For test-developers, these assessments provide improved test reliability, test validity, and test security (Gershon, 2005; Patsula, 1999).

While CATs have many advantages that make them an ideal testing approach, they also have their disadvantages. These disadvantages can directly affect the test-developer. Since tests are not assembled prior to administration, developers cannot review the test form. Large item banks must be developed and maintained in order to ensure particular items are not over-used and that all content is equally represented (Gershon, 2005; Patsula, 1999; Wainer & Eignor, 2010). Also, the initial costs of CAT development can be quite high (Meijer & Nering, 1999).

A disadvantage that specifically affects examinees is that CAT designs do not allow for common test-taking strategies (Mead, 2006). In a CAT, examinees are unable to skip questions or review questions once answered as they can in a conventional test (Gershon, 2005; Patsula, 1999). This is a controversial topic in the area, as item review may allow greater satisfaction amongst test-takers but could result in less precise ability estimates. Thus, examinees support the inclusion of review whereas test developers do not. MSTs provide a solution for this issue.

MST designs re-introduce some of the testing strategies examinees use in conventional testing but are unavailable in CATs. Due to the modular design, examinees may review responses to items within a module, which may decrease stress and anxiety for examinees. This

advantage of MSTs is an ideal aspect of conventional tests that was lost in CAT designs and is one of the most important advantages of MSTs (Hendrickson, 2007; Patsula, 1999; Robin, Steffen, & Liang, 2014; Yan et al., 2014; Zwick & Bridgeman, 2014). Another strategy examinees can utilize is item skipping; they may skip items within a module that are too difficult and return to them if time permits (Zwick & Bridgeman, 2014).

MSTs also have several advantages over both conventional tests and CATs. Test-developers have more control over the test in terms of balancing content. Since modules are pre-assembled, this provides an opportunity for review before administration (Breithaupt, Ariel, & Veldkamp, 2005; Chuah et al., 2006; Hendrickson, 2007; Yan et al., 2014; Zenisky & Hambleton, 2014; Zwick & Bridgeman, 2014). This review can be item by item, but often, module item reports can be used; this is faster and troublesome items can be flagged for further review (Luecht & Nungester, 1998). Due to increased ability for item review prior to administration, item and test security is often higher with MSTs (Breithaupt et al., 2005; Hendrickson, 2007). Modular presentation of material also allows items to be scored as a unit with a polytomous IRT model and interdependency is not a concern. Lastly, since there is less adaptability in MSTs (i.e., less routing points) than a CAT, test administration (e.g., scoring, routing, data management, computer processing) is more efficient (Hendrickson, 2007).

Although MSTs have some of the same advantages as CATs and do offer other advantages, they do have disadvantages. In order to create parallel modules and panels, large item pools are required (Breithaupt et al., 2005; Hendrickson, 2007). While there are many advantages for item writers and item developers in terms of review and control, this does mean an increase in work and time needed that CATs do not require. MSTs have the potential for increased error if only two stages are used, as one routing point might not be sufficient for

measurement precision (Hendrickson, 2007). Lastly, to approximate the results of a CAT, more stages and modules at each stage are necessary, which increases the complexity of the MST (Patsula, 1999).

Subgroup Differences

Assessments use various techniques in order to ensure the test is fair and unbiased against different population subgroups. However, while a test may be fair and unbiased, it does not guarantee group differences in cognitive abilities, represented by test performance, will not exist. For example, examination of scores by military recruits on the Armed Services Vocational Aptitude Battery (ASVAB), as well as the Armed Forces Qualification Test (AFQT) composed of four ASVAB subtests, show that women and racial minority group members often score lower on these tests (GAO, 1990). The scores influence the selection of these members and placement within the military, as they are less successful predictors for women and minority groups than white males. Thus, the GAO (1990) called for an examination of the sensitivity and fairness of these tests. Wise et al. (1992) examined the ASVAB and its technical composites. They concluded that the test was unbiased and fair for women and African Americans, even though subgroup differences exist.

Differences in cognitive abilities between demographic groups, such as gender, ethnicity, and race, have been documented in the literature (Eitelberg, 1981; Lynn & Kanazawa, 2011; Roth, Bevier, Bobko, Switzer, & Tyler, 2001). Representation of these cognitive abilities may focus on different aspects. One approach is to examine general intelligence (*g*; Spearman, 1904, 1927), which suggests that all cognitive abilities are related to one common factor. However, this common factor has been further broken down. *G* can be further divided into two components: fluid (*Gf*) and crystallized (*Gc*) intelligence. These two categories suggest that

aspects of intelligence are innate (i.e., *Gf*) while other aspects are learned (i.e., *Gc*; Cattell, 1963; Horn & Cattell, 1966). Intelligence can also be represented as general influences, such as visualization, fluency, speediness, memory and learning, audition, and retrieval (Carroll, 1993; Horn & Cattell, 1966). These cognitive abilities are considered the second level of intelligence, underneath *g*, and can be further divided into more specific processes and abilities.

Roth and colleagues (2001) documented differences between Caucasians, African Americans, and Hispanics in performance on educational, military, and personnel selection assessments. Examination of these differences is important in relation to policy decisions and may explain wage differences (Blackburn, 2004). Using the *d* statistic, a standardized statistic representing the mean difference between two groups divided by the sample-weighted average of the group standard deviations, the authors conducted a meta-analysis supporting the existence of differences between ethnic groups on various tests. These tests represented general intelligence (*g*) and more specific abilities (i.e., verbal and mathematical ability) often measured via achievement tests, which have high correlations with intelligence tests. The researchers also looked at differences between ethnicities based on various moderator variables; this approach is important because, although differences may exist, it is often difficult to explain the differences since the groups often vary on many different levels.

The researchers found overall differences in *g*, ignoring all moderator variables, between Caucasians and African Americans ($d = 1.10$) and Caucasians and Hispanics ($d = .72$). This implies that, in both comparisons, Caucasians score higher on intelligence than the other two ethnicities. However, other variables did moderate the size of the standardized difference. For example, Caucasians perform higher than African Americans on military assessments (Eitleberg, 1981; Roth et al., 2011). The military status of an individual (i.e., applicant versus incumbent)

also influenced the size of the difference observed. Applicants had a *d* of 1.19 whereas incumbents had a lower *d* (.46) when comparing Caucasians to African Americans. Thus, differences do exist between ethnicities, but the size of these differences are also related to various moderators (Roth et al., 2011).

Lynn (2006) provides a comprehensive summary of race differences in various intelligence measures. The book finds differences in intelligence for ten different races. IQ scores range from approximately 60 to 100 points. Africans have a weighted IQ average of 67, Caucasians have a weighted IQ average of 99, and Asians have a weighted average of 103. As can be seen, races do vary in their IQ. Similar to Roth and colleagues (2011), standard deviation units for IQ are reported. Caucasians are used as the reference group. For Africans, the standard deviation unit is -2 for IQ (Lynn, 2006). Thus, this work provides more support to the notion that different races vary in IQ, as well as mathematics and science scores.

While a majority of studies find little to no group differences in test performance between genders (e.g., males and females), some studies do suggest that gender differences on specific tasks may exist. Males often exceed at visual-spatial and mathematical tasks whereas females perform better on verbal tasks (Eitelberg, 1981). These differences can also change over the course of time. Females have been found to have higher IQs at earlier ages (i.e., ages 7 and 11) but this advantage changes at the age of 16, when males have higher IQs (Lynn & Kanazawa, 2011). At this age, the difference was approximately 1.8 IQ points, or .12*d*, where *d* represents standard deviation units. Boys also had higher standard deviations than girls, suggesting a greater variance in intelligence. While differences in IQ were observed, the study only used children in Britain, and thus might not generalize to other populations.

Current assessments show differences in the ability of examinees based on their gender and ethnicity. The Graduate Record Examination (GRE) reported gender differences in Verbal and Quantitative scores for U.S. citizens who took the GRE between 1 July 2014 and 30 June 2015 (ETS, 2016). On average, men scored higher than women in both the verbal reasoning (approximately 3 points) and quantitative reasoning (approximately 4 points) sections. Differences in mean and standard deviation can also be observed between ethnic groups as well on the GRE. For example, African Americans tend to score lower on both assessments than Caucasians. Asians tend to score the highest on the Quantitative section.

Gender and ethnic differences have also been reported in SAT scores in college-bound seniors in 2015 in both mathematics and critical reading (i.e., verbal). When examining gender, males often score higher on both critical reading (4 points) and mathematics (31 points) than females (SAT, 2015). These distributions have different means and standard deviations, which might provide different information if applied during trait estimation. Differences in means and standard deviations also exist for various ethnic groups in mathematics and critical reading. For example, Caucasians have the highest mean score for critical reading and African Americans have the lowest (98-point difference). For mathematics, Asians have the highest score, whereas African Americans have the lowest (170-point difference; SAT, 2015).

Collateral Information

Additional information, termed collateral information, can be obtained for both items and people. For example, collateral information related to items may be features of the items (e.g., length, type of item, etc) such as those used in structural IRT models to predict item difficulty (Mislevy, 1988; Veldkamp & Matteucci, 2013). For people, or examinees, collateral information may be demographic variables (e.g., gender, race), socioeconomic status, age, grade, country, or

previous test scores. This type of information is contrasted with historical data, which is "data arising from previous similar studies where the same response variable and covariates of the current study have been collected" (Matteucci & Veldkamp, 2015, p. 919). Information such as this might be used to estimate both item and person parameters (e.g., difficulty, discrimination, latent ability). Therefore, the use and implications of collateral information in estimation must be examined.

Item Parameter Estimation

Studies have been conducted using collateral information to estimate item parameters. Mislevy and Sheehan (1989) used collateral information concerning examinees (e.g., age, grade) to estimate item parameters. The researchers examined multiple cases of collateral information; they examined situations in which no collateral information was known, collateral information was known but not used, collateral information was known and used in examinee sampling, and collateral information was known and used in both examinee and item sampling. The main conclusion of the study was that, if collateral information concerning examinees was used in order to assign items (i.e., only items appropriate for a specific grade were assigned), then the collateral information (i.e., grade) should be used when estimating item parameters. Otherwise, parameter estimates would be inconsistent.

Another study examined the utility of using an empirical prior distribution versus non-empirical priors in estimating items parameters (Matteucci, Mignani, & Veldkamp, 2012). The study examined the possibility of having item covariates for item discrimination and item difficulty estimation. Results support the use of an empirical prior over a non-empirical prior. Bias in both discrimination and difficulty was reduced using the empirical prior, particularly in the extreme regions of the item difficulty continuum. However, the researchers only examined a

fixed-length linear test, and thus, the approach should be applied to an adaptive context before being used in such scenarios. Regression trees have shown to have similar results for item parameters when creating an empirical prior from features of the items (Veldkamp & Matteucci, 2013) or the examinees (Matteucci & Veldkamp, 2011) by examining the regression relationship between the covariate(s) and ability being examined. Response times have also increased the efficiency of item parameter estimation (van der Linden, Entink, & Fox, 2010).

Ability Estimation

While the use of empirical information (i.e., collateral information) in item parameter estimation is less controversial, ethical issues exist related to the application of empirical priors in ability estimation (Veldkamp & Matteucci, 2013). In certain situations, such as high-stakes achievement and aptitude testing, utilization of an empirical prior might have disadvantageous effects, such as bias in ability estimation based on group membership (i.e., implicit stereotypes). To circumvent these issues, researchers and test developers/administrators suggest the use of empirical information during test start-up and administration, such as when selecting the next item in a CAT, but to simply use the response patterns to estimate the examinee's final ability (Matteucci & Veldkamp, 2013; van der Linden, Entink, & Fox, 2010; Veldkamp & Matteucci, 2013). Research on the application of collateral information in ability estimation has persevered.

To exploit collateral information in ability estimation, a relationship between ability and the collateral information (e.g., covariates) must be created (Matteucci & Veldkamp, 2011; van der Linden, 1999). The relationship between θ_i, representing the latent ability of examinee *i*, and X_P, representing the set of *P* individual covariates in the set [$X_P = 1, \ldots P$], is represented by Equation 2.18,

$$\theta_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{iP} + \epsilon_i \,, \qquad 2.18$$

where the β's represent coefficients and the error terms are assumed to be independent and identically distributed ($\epsilon_i \sim N(0, \sigma^2)$). This linear regression is translated into a normal conditional distribution of θ_i given the set of X_{iP} covariates:

$$\theta_i | X_{i1}, \ldots, X_{iP} \sim N(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{iP};\ \sigma^2)\ . \qquad 2.19$$

Equation 2.19 represents the use of an empirical prior distribution for ability when examinees are randomly sampled from a subpopulation within X_P (van der Linden, 1999). Empirical information may also be used to calculate an initial ability estimate in order to initialize ability estimation in an adaptive design and select the first item administered using Equation 2.20:

$$\widehat{\theta_i} = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{iP}\ . \qquad 2.20$$

This approach provides a better provisional estimate for an individual, which might decrease the time needed to converge on an ability estimate, as well as leads to higher statistical precision and lower item exposure (Matteucci & Veldkamp, 2013; van der Linden, 1999; Veldkamp & Matteucci, 2013).

Various studies have utilized this approach to empirical priors for estimating ability. Through direct estimation of the regression weights and the variance of the prior, σ^2, van der Linden (1999) crated a relationship between the Name Comparison test and Vocabulary test included within the adaptive version of the Dutch General Aptitude Test Battery. Previous performance on one test (i.e., Name Comparison) was used as collateral information for ability initialization or ability estimation using an empirical prior. Veldkamp and Matteucci (2013) used performance on one construct (e.g., intelligence via Raven's Matrices) to create an informed, empirical prior for performance on a similar construct (e.g., intelligence via Number Serires). Using a CAT simulation design, the researchers found that for θs close to 0, slightly shorter tests could be administered to examinees. For more extreme θs (i.e., high and low ends

of the continuum), a considerable reduction in test length was observed when using empirical priors.

Matteucci and Veldkamp (2013) also examined the use of empirical priors in a CAT design, comparing fixed-length and variable-length tests. Three simulation designs were examined: one simulation was *standard*, using a standard normal distribution as a prior and ability is initialized at a point estimate of zero; one simulation used only *empirical initialization*, in which an empirical prior was used as an initial estimate; and the last simulation was *fully empirical*, utilizing an empirical prior for initialization and estimation. Results show that the fully empirical situation reduced the number of items needed in relation to the standard simulation, also reducing item exposure. This situation also resulted in more precision ability estimation over the standard situation in both variable-length and fixed-length CATs. The empirical initialization situation provided intermediate results (i.e., its behavior was between the other two situations). Empirical initialization performed better than the standard situation, but was less precise than the fully empirical situation. The researchers also conducted two empirical studies, one in an intelligence test setting and the other in an educational setting (Matteucci & Veldkamp, 2013). Both of these empirical situations used performance on a previous test as the collateral information for the succeeding test. In both studies, the fully empirical design performed the best, especially at the extreme ends of the ability continuum.

The studies reported above often used MCMC with a Gibbs sampler to estimate the marginal maximal likelihood (MML) estimation of ability. The studies also suggest that measurement precision and [reduced] bias will be related to the quality of the collateral information. However, all of the studies have used performance on one test as collateral information. None of the studies directly examined the impact of priors based on group

membership for the reasons stated above. While researchers suggest the use of these demographic empirical priors for initialization and item selection only, no studies have been conducted examining the applicability of these approaches.

CHAPTER 3

DATA AND METHODOLOGY

This chapter provides information pertaining to the design of the simulation study, including data generation, design conditions, and data analyses. The chapter is split into five sections: data generation, simulation test designs, ability estimation approaches, data analysis, and computer programs.

The main goal of the study is to examine the influence of various group-based test priors on trait estimation across the ability continuum in assessments, particularly CAT designs. It is hypothesized that the use of inappropriate versus appropriate priors, based on group membership, will differentially impact estimates across groups. Standard error (SE), bias, and Root Mean Square Error (RMSE) will be examined for different populations at the total population level, group membership level, and theta score level. Each of these statistics will be examined for two fixed-length tests (CTs) and two CAT designs. Based on current knowledge, it is expected that CAT designs will have lower standard errors (i.e., more precise measurement), less bias, and higher accuracy (i.e., lower RMSE values) when compared to the fixed-length test. These routine expectations arise since CAT designs tailor the assessment to the individual, removing the ability-difficulty mismatch common in fixed-length tests. Two CAT simulations will be conducted, in which length is varied. Again, based on current information, it is hypothesized that the longer CAT will have lower standard errors than the shorter assessment.

Bias, representing a tendency to either over- or underestimate ability on an assessment, will be examined. It is expected that bias will be present in the final estimates when using known information, in the form of an informative prior, about the simulee. In general, estimates will be pulled toward the mean of the prior distribution utilized. For simulees farther from this

mean, bias will be higher than for those closer. In other words, simulees' whose true ability is closer to the ability continuum extremes will be more heavily influenced by the prior, since it provides biased information pertaining to their true score (i.e., the information provided is that their true ability is closer aligned to the majority of the population when it is not). It is pertinent to look at various theta score levels, as bias might exist at the extremes but not observed in overall bias (i.e., one extreme is underestimated, the other is overestimated, and this effect cancels out when examining overall bias). However, this influence should be mitigated by the presence of information within the data (e.g., the scored responses). However, if there is not enough information within the data or the prior utilized is extremely poor, bias may still be high.

Greater bias should be observed overall and for theta score levels when an inappropriate prior is used instead of an appropriate one. An appropriate prior is one in which the prior's mean is equal to the mean observed in the population or population subgroups (i.e., true prior). An inappropriate prior is one where the means are different; the prior's mean may be above or below that observed in the total population or population subgroups. For example, if a prior is chosen with a mean based on the entire population but various subgroups exist whose true ability mean is different, estimates for these subgroups will be more biased than for those subgroups whose true mean is closer to that of the prior. In contrast, when an less informative prior (i.e., uniform prior) is used for final ability estimation, bias will be less extreme across the entire continuum but may still be higher than desired, since there is no capitalization on previous information pertaining to the individual.

Lastly, accuracy concerning the ability estimates will be examined via the use of RMSE. Smaller RMSE values indicate higher accuracy levels. Higher accuracy will occur when the prior aligns closer to the true distribution of ability in the population (i.e., an appropriate prior is

used). Thus, estimates will be less accurate when inappropriate prior distributions are utilized for ability estimation. Accuracy will suffer more towards the extremes of the ability continuum (i.e., these estimates are being pulled to the mean of the prior).

Data Generation

Item Parameters

Parameters were simulated to represent a four choice, multiple-choice assessment. Item parameters were generated as if calibrated using the Rasch (Equation 2.4) and 2PL (Equation 2.2) IRT models under the normal metric, resulting in two item banks. In the 2PL item bank, each item has an item difficulty and an item discrimination parameter. In the Rasch item bank, item discrimination was constrained to 1 with varying item difficulty parameters. While there was no set rule on the required item bank size, enough items were generated to accurately represent the latent continuum (e.g., appropriate difficulty) and provide enough information at different locations (Green et al., 1984; Weiss & Kingsbury, 1984). Items for both banks were generated along the ability spectrum between $\pm$ 3 at 0.25 increments (i.e., bins), with 100 items generated in each bin. This resulted in a total of 2,500 items per bank. Thus, item difficulties were normally distributed, $\beta \sim N(\mu_b, 0.04)$, such that μ_b represented the current location on the continuum. All item discriminations were generated from a log-normal distribution with a mean of 0 and a variance of 0.25 (i.e., $\alpha \sim logN(0, 0.25)$), mimicking the default parameters in BILOG-MG 3 (du Toit, 2003; Zimowski, Muraki, Mislevy, & Bock, 2003).

Table 3.1 presents that means and standard deviations for the 2,500 items in each item bank. The mean difficulty was 0, as expected. The item parameter distributions were examined graphically. As seen in Figure 3.1, the item difficulty distributions for both items banks (i.e., B and D) were approximately uniform, suggesting there was roughly the same number of items at

each position on the latent continuum. The items discriminations were equal (i.e., 1) for the Rasch item bank (i.e., A) and were clustered around a mean of 1 for the 2PL item bank (i.e., C)

Table 3. 1. *Means and standard deviations of the total item bank.*

IRT Model	*N*	Item Parameter	*M*	*SD*
Rasch Item Bank	2,500	α	1.0000	0
		β	0.0005	1.8148
2PL Item Bank	2,500	A	1.1469	0.6106
		β	-0.0069	1.8157

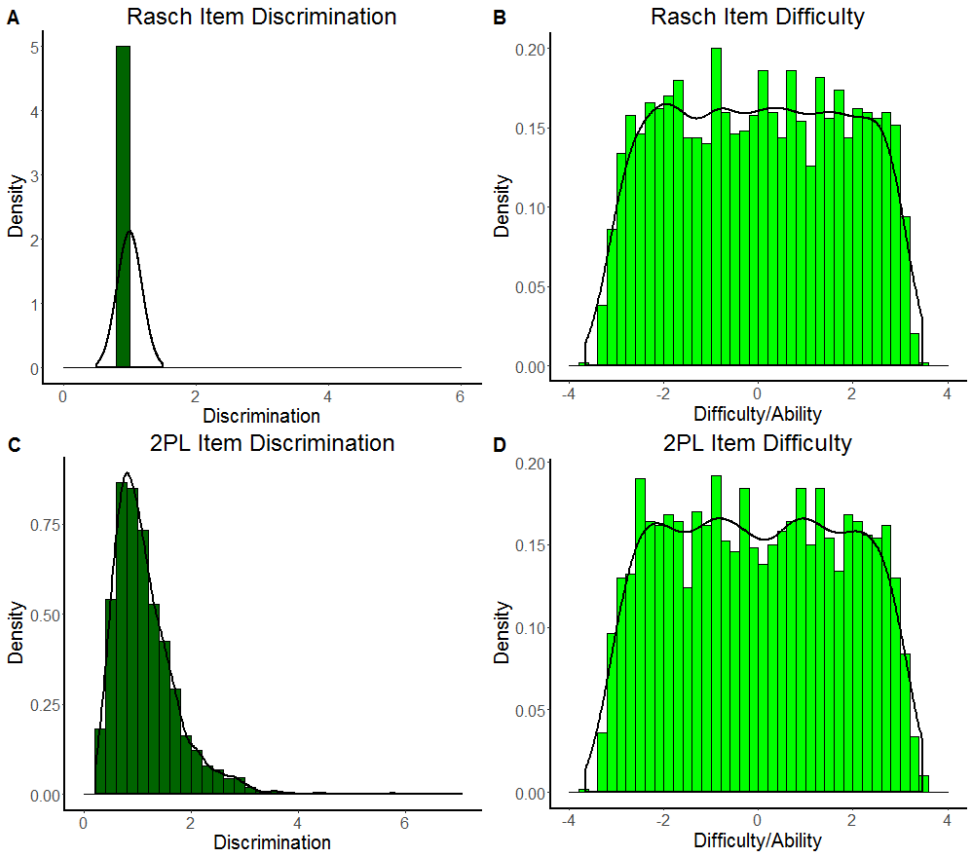

Figure 3. 1. Distribution of item parameters in the Rasch (A and B) and 2PL (C and D) item banks.

Person Parameters

Two populations were generated to represent true ability for two hypothetical groups, Group A and Group B. Group A represented the majority, whereas Group B represented the

minority. Each population contained 500,000 simulees; Group A represented 70% of the total population and Group B composed the remaining 30% of the population. This population composition was chosen to be similar to census data, which reports approximately a 75% majority (i.e., Caucasian) and 25% minority composition (U.S. Census Bureau, 2016). Table 3.2 presents the distributional information for the total population for each simulation. The Single Population (i.e., Simulation One) represented a situation in which both groups had the same true ability distributions; this represented a baseline condition. The Subgroup Population (i.e., Simulation Two) represented a situation in which the two groups differ in true ability distributions, each having a different mean. In this condition, the majority group (i.e., Group A) had a true ability distribution with a mean of 0.5 and a standard deviation of 1. The minority group was split into two subgroups. The first subgroup (i.e., Group B Low, or B_L) was simulated to have a mean a full standard deviation below the majority; it consisted of 80% of the minority group. Thus, its mean was -0.5 and its standard deviation is 1. The second subgroup (i.e., Group B High, or B_H) was generated to have the same true ability distribution as the majority group and contained the remaining 20% of the minority group.

Table 3. 2. *True ability distributions for the populations.*

Simulation	Population	Group	N	Proportion	True Ability
Simulation One	Single Population	Group A	350,000	0.70	$\theta_{TA1} \sim N(0,1)$
		Group B	150,000	0.30	$\theta_{TB1} \sim N(0,1)$
Simulation Two	Subgroup Population	Group A	350,000	0.70	$\theta_{TA2} \sim N(0.5,1)$
		Group B_L	120,000	0.24	$\theta_{TB2.L} \sim N(-0.5,1)$
		Group B_H	30,000	0.06	$\theta_{TB2.U} \sim N(0.5,1)$

Descriptive information, examined after the person data were generated, supported the intended pattern from Table 3.2. Table 3.3 shows that the mean for both groups in the Single

Population were approximately 0 (Figure 3.1A). For the Subgroup Population, the population composite mean was 0.26 and the standard deviation was approximately 1. The means were also obtained for the groups for in the Subgroup Population. Group A had a mean of 0.5 and Group B had a composite mean of -0.3 (Figure 3.1B). The individual group means for Group A, Group B_L, and Group B_H were also obtained (Figure 3.2).

Table 3. 3. *Descriptive information for the true ability conditions.*

Simulation	Condition	Group	N	Mean	SD	Min	Max
Simulation One	Condition 1	A	350,000	-0.0032	0.9988	-4.9286	4.5781
		B	150,000	0.0012	1.0042	-4.2161	4.4927
		Total	500,000	-0.0018	1.0005	-4.9286	4.5781
Simulation Two	Condition 2 Composite	A	350,000	0.5000	1.0005	-3.9506	5.6225
		B	150,000	-0.2988	1.0727	-4.5362	4.8828
		Total	500,000	0.2604	1.0862	-4.5362	5.6225
	Condition 2 Specific	A	350,000	0.5000	1.0005	-3.9506	5.6225
		B_L	120,000	-0.4984	0.9968	-4.5362	3.7960
		B_H	30,000	0.4996	0.9911	-3.7955	4.8828

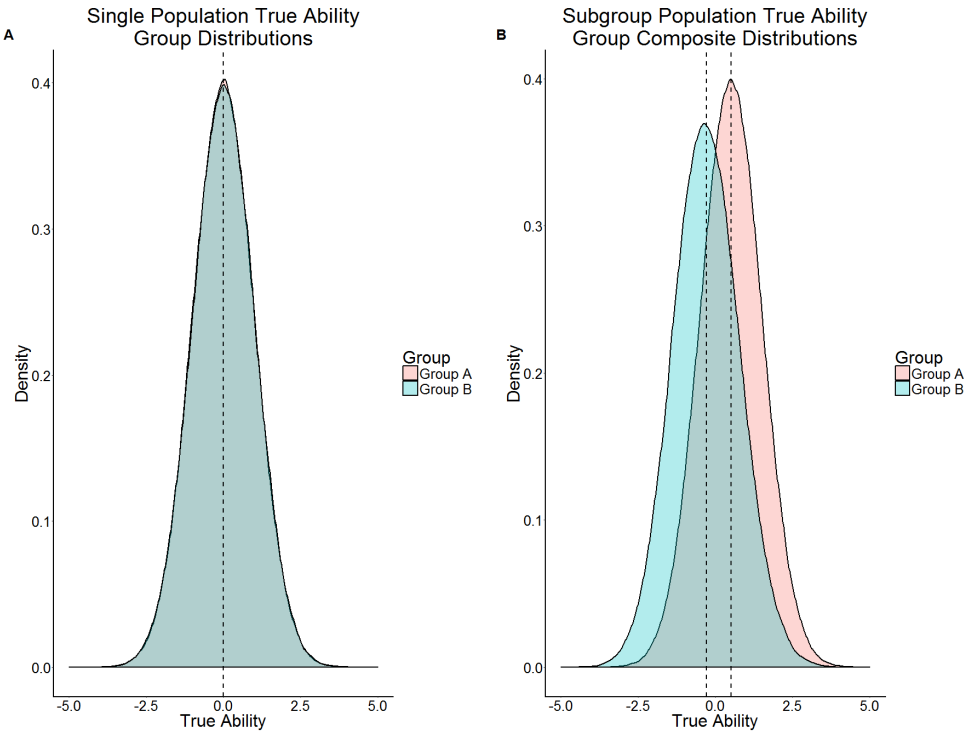

Figure 3. 2. True ability distributions by group for Single (A) and Subgroup (B) Populations.

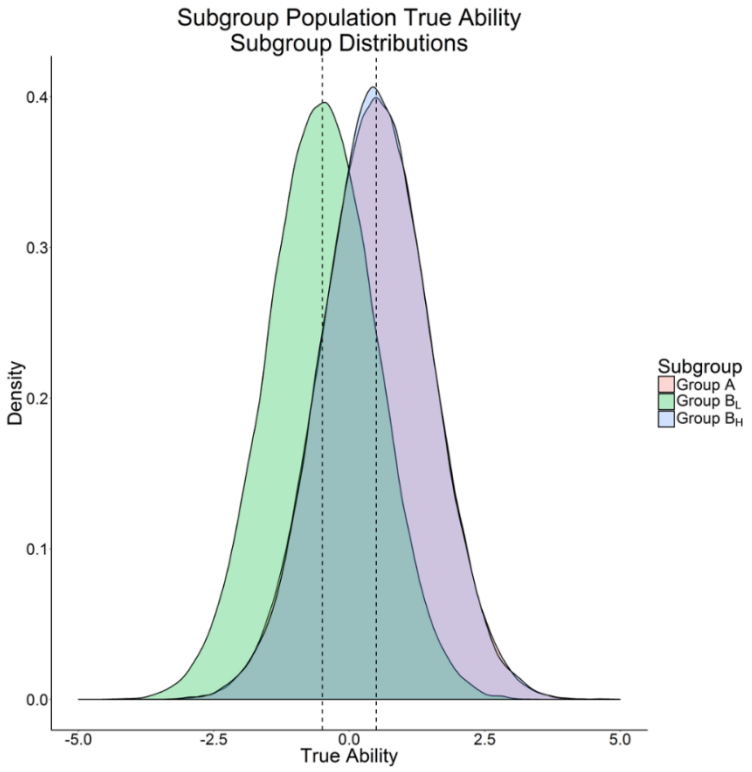

Figure 3. 3. *True ability distributions for the Subgroup Population subgroups.*

Simulees were classified into one of ten theta score levels based on their true ability. Table 3.4 shows the theta groupings for each score level. The number of simulees in each theta score level for each population is presented in Table 3.5. While simulees do exist in every theta score level, it is possible to have a percentage of 0 when compared to the total population.

Table 3. 4. *Theta score levels for comparisons across the ability continuum.*

Theta Score Level	Theta Ranges	Theta Score Level Mean
1	$\theta < -2$	$\theta_1 < -2$
2	$-2 \leq \theta < -1.5$	$\theta_{2Mn} = -1.75$
3	$-1.5 \leq \theta < -1$	$\theta_{3Mn} = -1.25$
4	$-1 \leq \theta < -0.5$	$\theta_{4Mn} = -0.75$
5	$-0.5 \leq \theta < 0$	$\theta_{5Mn} = -0.25$
6	$0 \leq \theta < 0.5$	$\theta_{6Mn} = 0.25$
7	$0.5 \leq \theta < 1$	$\theta_{7Mn} = 0.75$
8	$1 \leq \theta < 1.5$	$\theta_{8Mn} = 1.25$
9	$1.5 \leq \theta < 2$	$\theta_{9n} = 1.75$
10	$2 \leq \theta$	$2 \leq \theta_{10}$

Table 3. 5. *Theta score level counts, and population percentage, for the total population.*

Theta Score Level	Single Population		Subgroup Population			
	Group A	Group B	Group A	Group B	Group B_L	Group B_H
1	7949 (2%)	3527 (1%)	2144 (0%)	8096 (2%)	7900 (2%)	196 (0%)
2	15489 (3%)	6719 (1%)	5823 (1%)	11308 (2%)	10843 (2%)	465 (0%)
3	32466 (6%)	13666 (3%)	15419 (3%)	19532 (4%)	18278 (4%)	1254 (0%)
4	52459 (10%)	22332 (4%)	32248 (6%)	25647 (5%)	22909 (5%)	2738 (1%)
5	67042 (13%)	28573 (6%)	52438 (10%)	27476 (5%)	22957 (5%)	4519 (1%)
6	67143 (13%)	28760 (6%)	66934 (13%)	23931 (5%)	18057 (4%)	5874 (1%)
7	52177 (10%)	22537 (5%)	66954 (13%)	16947 (3%)	11128 (2%)	5819 (1%)
8	32043 (6%)	13672 (3%)	52627 (11%)	9718 (2%)	5237 (1%)	4481 (1%)
9	15371 (3%)	6722 (1%)	31813 (6%)	4629 (1%)	1956 (0%)	2673 (1%)
10	7861 (2%)	3492 (1%)	23600 (5%)	2716 (1%)	735 (0%)	1981 (0%)

Simulation Test Designs

Two testing designs were examined for each population: conventional, fixed-length tests (CT) and computer adaptive testing (CAT). Each test underwent 100 replications to test consistency. A random sample of 3,000 simulees was drawn from the total population, specified in Table 3.2, for each replication. Simulees in the sample were divided into 10 theta score levels based on true ability; each of these groups were equally represented in the sample. This stratification was done to ensure that results were not influenced by outliers or too little information in a level. For Simulation One (Single Population), a total of 1,500 simulees were pulled from each group (i.e., Group A, Group B); 150 simulees were sampled from each of the theta score levels. For Simulation Two (Subgroup Population), a total of 1,000 simulees from each group (i.e., Group A, Group B_L, Group B_H) were sampled; 100 simulees were sampled from the theta score levels to have equal representation in all subgroups. Thus, the subgroups will not differ in their overall score means, as observed in the population. This ensured that each group, as well as each theta score level, had an equal representation of simulees in each simulation. However, the sample will not have a distribution like the total population due to this design. To combat this issue in the analyses, a subset of the simulation samples will be drawn to approximate the individual subgroup distributions in the population. Both types of data will be examined.

Conventional Test

First, simulees were routed through two conventional (i.e., fixed) testing designs. These tests varied in length; one was composed of 15 items while the other was composed of 30 items. For each test, each simulee was administered the same set of items, in the same order, to obtain $\hat{\theta}$. Four conventional tests, utilizing a specific item bank, were created for each population (i.e.,

eight tests total) to have peaked item difficulties at the mean of the composite population distribution (i.e., total population mean). This approach aims to have high measurement precision and low standard errors of measurement (Mead & Drasgow, 1993; Weiss, 1985, 2011). Thus, the items were clustered around this mean instead of spanning the whole item difficulty continuum. All items were selected from the item bank to have high levels of item discrimination ($\alpha_i \geq 1.5$) for the 2PL IRT model.

Table 3.6 provides descriptive information for the conventional tests under each IRT model and test length. For the Single Population, both tests were constructed to have means of 0, while the items for the Subgroup Population were chosen such that the test's mean difficulty was 0.26. For the 2PL model, item discriminations were, on average, greater than 1. Appendices A and B contain the item parameters for the conventional tests for the Single and Subgroup Populations, respectively.

Table 3. 6. *Descriptive statistics for all conventional test designs for the two simulations.*

Simulation	Population	IRT Model	N	Item Parameter	M	SD
Simulation One	Single Population	Rasch	15	α	1.0000	0.0000
				β	0.0332	1.1650
			30	α	1.0000	0.0000
				β	-0.0652	1.1064
		2PL	15	α	1.8916	0.4507
				β	0.0062	1.1629
			30	α	2.1067	0.6127
				β	-0.0112	1.1059
Simulation Two	Subgroup Population	Rasch	15	α	1.0000	0.0000
				β	0.2527	0.4479
			30	α	1.0000	0.0000
				β	0.2666	0.7894
		2PL	15	α	1.9949	0.3859
				β	0.2664	0.6265
			30	α	2.3002	0.7532
				β	0.2550	0.7207

CAT

Two separate CAT simulations were conducted to represent fixed-length CAT design, like the conventional tests (Gershon, 2005). However, each successive item was selected based on the simulee's current $\hat{\theta}$ and the item's level of information. This approach, maximum information, chooses the most informative item at the simulee's current ability location using Fischer's information (Equation 2.5). Often, this is a popular choice in CAT administrations (van der Linden & Pashley, 2010). Two fixed-length CATs were conducted, one with a length of 15 items and the other with a length of 30 items. To avoid any political and legal issues, only fixed-length designs were used.

For each replication, a different item pool, or a subset of items from the total item bank, was used to simulate real-world applications. A total of 100 item pools was used (1 per replication) for all various conditions. This approach was utilized because practical applications of CAT assessments do not have an infinite number of items at each level during administration. These item pools were created by selecting six items from each of the 25 bins. For the Rasch model, six items were randomly selected. For the 2PL model, one item was selected such that $\alpha_i < 0.5$, four items were selected such that $0.5 \leq \alpha_i \leq 1.5$, and the last item was selected so that $\alpha_i > 1.5$. This was to simulate an item pool containing both high and low discriminating items; otherwise, the item pool might have had all highly discriminating items.

Each CAT administration started with the most informative item (i.e., using maximum information item selection) at the mean of the chosen test prior. For example, for a test prior based on the total population, each simulee was first administered an item around this difficulty. Therefore, each simulee received the same item at the beginning of the test. However, when group specific priors were used, the starting item varied for each simulee based on group

membership. While this approach did result in the same few items frequently used at the beginning of the test, this was not a cause for concern in the simulations. Item exposure was not examined in the study and thus not controlled. Also, CAT assessments often start with the same item to ensure that all examinees have the same starting point.

Ability Estimation

Trait Estimation

EAP (Equation 2.12; Bock & Aitkin, 1981) estimation was used to estimate ability. It is less computationally intense, less biased, estimates proficiencies for people at the extremes, and can be used with perfect response patterns, which makes it preferred over other approaches, such as MLE. Using EAP estimation allowed for the presence of information about the person/population via the inclusion of a prior distribution. In this study, EAP with a specific, informative prior based on population groups was used to initiate the test (i.e., initial item selection) and during test administration (e.g., item selection after each successive item response) for CATs, but only influenced CTs at the end of testing. However, final estimates were obtained using two approaches.

Two different methods to final ability estimation were used. Method 1 (i.e., EAP Normal) used EAP estimation to obtain the simulee's final ability estimate in conjunction with the informative test prior, discussed next, utilized during test administration. In other words, the test prior was used to start the test via ability initialization at the prior's mean, item selection during the assessment, and for final ability estimation. The EAP estimate used 40 equally-spaced quadrature nodes between $^{+}/_{-}$ 4 standard deviations from the mean of the informative prior. Weights were normally distributed, in which extreme nodes have smaller probabilities. Method 2 (i.e., EAP Uniform) used the approach from Method 1 (e.g., the informative prior is

used during all stages of test administration) with the exception that the final ability estimate was estimated using a less informative prior, which provided little influential information about the simulee and used primarily the examinee's scored responses to obtain a final estimate. While the score responses provide a majority of the information, there is some impact from the uniform prior. As with the EAP Normal estimate, the EAP Uniform estimate used 40 equally-spaced quadrature nodes between $^{+}/_{-}$ 4 standard deviations from the mean of the informative prior. However, all weights were equal; in other words, each quadrature node had a similar probability. This second method simulated the approach of using empirical information to start and administer the assessment but using predominantly the response patterns to estimate the simulee's final ability estimate, as recommended by researchers and test developers.

Prior Information Utilization

Three test prior scenarios were employed to examine the impact of different test priors on trait estimation under the true ability distributions and trait estimation approaches. Table 3.7 represents the test prior distribution scenarios utilized. The hyperparameters for each prior are fixed, not estimated. Scenario 1 (i.e., Population Composite Prior) represented the use of the composite population mean obtained from the entire population. This scenario was utilized in both simulations. As can be seen, for the Single Population, the population composite mean is 0; for the Subgroup Population, the population composite mean is 0.26.

The last two scenarios were only used with the Subgroup Population. Scenario 2 (i.e., Group Composite Priors) represented the use of group composite priors based on group membership (X_{gc}). Thus, for Group B, the composite prior was composed of the two subgroups (i.e., mean of both the Group B low and high subgroups). For an individual in Group A, X_{gc} was 0 and a mean of 0.5 was used. However, for an individual in Group B, X_{gc} was 1 and a mean of -

0.3 was used. Scenario 3 (i.e., Group Specific Priors) was like Scenario 2, except members of Group B were further divided into the high and low subgroups. Thus, three individual priors were utilized. For Group A, the mean was 0.5. For Group B, the covariate X_{gs} represents the specific subgroup membership. When X_{gs} was 0, the mean was -0.5 to represent the lower subgroup (B_L); when X_{gs} was 1, the mean was 0.5 to represent the higher subgroup (B_H). While previous work has utilized a previous test score as a prior (van der Linden, 1999; Veldkamp & Matteucci, 2013), the current work did not since the focus was on how composite group priors, created via the use of the total population or group's mean, impacted trait estimation and not how individualized priors impact estimation. Therefore, the pertinent collateral information in this simulation is group membership.

Table 3. 7. *Prior distribution scenarios employed in the simulation.*

Simulation	Population	Scenario	Label	Prior Distribution
Simulation One	Single Population	Scenario 1	Population Composite Prior	$\theta_{1.P1} \sim N(0,1)$
Simulation Two	Subgroup Population	Scenario 1	Population Composite Prior	$\theta_{2.P1} \sim N(0.26,1)$
		Scenario 2	Group Composite Priors	$\theta_{2.P2} \sim N(-0.8X_{gc} + 0.5, 1)$
		Scenario 3	Group Specific Priors	$\theta_{2.P3A} \sim N(0.5, 1)$ $\theta_{2.P3B} \sim N(X_{gs} - 0.5, 1)$

Priors will be utilized different for CTs and CATs. For CTs, the above test priors will only be influential at the end of the assessment when estimating final ability. Therefore, these priors are considered final estimate priors. However, for the CATs, test priors are utilized throughout the entire testing process (administration and final ability estimation). Thus, they are called test priors.

Data Analysis

Correlations will be computed between the true ability parameters and the ability estimates across all replications from each ability estimation approach. High correlations indicate that the estimation approach produced estimated abilities like the true values. In conjunction to these correlations, three statistics will be used as dependent variables to examine the data in terms of θ and $\hat{\theta}$. First, precision of $\hat{\theta}$ will be examined via calculation of the standard error of the estimate (Equation 2.13). A mean standard error for data will be examined, and will be calculated using Equation 3.1.

$$\overline{SE_{\hat{\theta}}} = \sqrt{\frac{\sum_{i=1}^{N} SE_i^2}{N}} \qquad 3.1$$

The lower the standard error, the more precise the estimate. Second, residuals, representing bias, will be calculated for each simulee by subtracting his or her true ability from the estimated ability, $r_i = \widehat{\theta_i} - \theta_i$. These residuals represent the amount of error in the estimate and are expected to be 0. Using these residuals, the mean bias can be calculated using Equation 3.2 (Gorin et al., 2005; Patsula, 1999), below:

$$\bar{r} = \frac{\sum_{i=1}^{N} r_i}{N} \qquad 3.2$$

When this mean is non-zero, the ability estimates are considered to be biased and can be in either a positive or negative direction. Thus, in terms of this study, the mean bias will imply whether the approach to ability estimation reflects the tendency to over- or underestimate ability. Positive bias indicates an overestimation of ability, while negative bias indicates an underestimation of ability. Third, accuracy will be examined via the use of Root Mean Square Error (RMSE), which is the square root of the Mean Square Error (MSE) and is calculated using the below equation.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} r_i^2}{N}} \qquad 3.3$$

RMSE is expected to be 0 when accurate ability estimates exist.

For each of these statistics, values were analyzed in multiple. Analyses were constrained to examination of the three subgroups (A, B_H, B_L) for all combination of IRT model, test type, test length, test prior distribution, and estimation approach. These statistics were examined when all theta score levels are equally represented and when the groups approximate the population distribution. When examining the statistics for equal theta score levels, how the test priors affect ability across the entire ability continuum can be examined. For instance, does a test prior differentially affect different theta levels. Then, examining the statistics with a sample similar to the population allows for generalization of the results. However, it is inappropriate to examine group membership and theta score level simultaneously, and these analyses will be done separately.

Analysis of variances (ANOVAs) will be conducted using SE, bias, and RMSE as the dependent variable for each condition in both simulation studies. Under each IRT model (i.e., Rasch and 2PL), the independent variables that will be examined will be the assessment type, final trait estimation approach, and test prior. These ANOVAs will show any significant differences between the various conditions and their interactions.

Software

Several software programs were utilized in completion of the study. R (R Core Team, 2015) was used for multiple purposes. This software was used to create item pool and person samples, analyze data, and create tables and figures. Simulated data was analyzed using SAS/STAT software® for final ability estimation under both estimation methods (i.e., EAP

Normal and EAP Uniform). Two other software programs were utilized to obtain simulated data for all examinees.

SimulCAT (Han, 2012) is a simulation software that enables CAT simulations to be conducted. The software enables the user to put specify different conditions (e.g., prior, estimation technique, etc.) on the simulation design. The software reads in a person dataset, containing ID and true ability, as well as an item parameter file. Currently, unidimensional dichotomous models are available. Then, the user sets information concerning the items, such as: item selection criterion; test termination criterion; item exposure control; and content balancing. In this simulation, maximum information (MFI) is used to select items, a fixed-length CAT termination criterion is used in which the length of the test is varied (e.g., 15 and 30 items), and there is no item exposure control or content balancing in place. Next, information pertaining to test administration must be specified, like: approaches to obtaining initial, interitem, and final ability estimates; pretest items; replication datasets; seed values; and output to save. For the current study, the initial item is selected using a fixed value (e.g., the mean of the prior) and EAP estimation is used for all ability estimates. The mean and standard deviation of the prior can be set by the user; SimulCAT uses 40 equally-spaced quadrature points ranging from $^{+}/_{-}$ 4 standard deviations from the mean of the prior.

Once all this information is specified, the software conducts the simulation. The first item is administered by selecting the item that provides the most information at the mean of the chosen prior, which is the first provisional estimate given to the individual. This results in the same first item being administered to all examinees. Using the item's parameters and the individual's true ability, a probability is generating under the chosen IRT model. This probability is compared to a random number between 0 and 1. If the probability is greater than

the random number, the item is scored as correct (i.e., 1); otherwise, the item is scored as incorrect (i.e., 0). Using this response, the individual's estimated theta is updated. Using this updated theta, the most informational item is drawn from the item pool and administered. This process continues until the termination criterion is satisfied (i.e., the pre-determined number of items has been administered), at which point the final estimate of ability is obtained.

MSTGen (Han, 2013) is a MST simulation software that allows the user to place items into modules spaced across various stages. While MST simulations will not be conducted in this simulation, MSTGen will be employed to run the conventional test simulations. A conventional test is essentially a one-stage MST design with one module in the first stage. Therefore, MSTGen is useful for simulating the conventional tests runs. Set-up is similar to SimulCAT in terms of people and items. However, in MSTGen, the modules must be compiled. For this simulation, one module is created in which all the conventional test items are administered in a specific order. Ability estimation issues are handled in the same fashion; the mean of the prior is the initial person estimate. For MSTGen, though, the first item is administered regardless of the initial estimate or informational value. A probability is generated and a scored response provided as described above. Each item in the module is administered in the same order to all simulees. The final ability estimate is obtained using all item parameters and scored responses.

Summary

Table 3.8 presents a summary of the study design. The table is divided by the two populations: Single Population (Simulation One) and Subgroup Population (Simulation Two). It is also divided by the test type in Simulation Two, since CATs and CTs will be run separately due to the number of priors involved. The number of priors did not vary for CTs and CATs in Simulation One, so separate runs were not necessary. There are three within-subject variables

being examined throughout the analyses: group membership (Table 3.2), theta score level (Table 3.4), and final trait estimate type (2). There are three between-subject conditions being examined: IRT Model (2), Test Length (2), and Test Prior. However, the IRT models will be analyzed separately. The two simulations differ in the number of test priors. The Single Population only has one test prior scenario with 8 conditions (i.e., Testing Type x Test Length x Prior During Test). The Subgroup Population, in contrast, has more test prior scenarios, resulting in 12 total CAT conditions (i.e., Testing Length x Final Trait Estimation x Prior During Testing). However, for the CT, there are 8 conditions (i.e., Test Length x Final Estimate Priors). The CTs and CATs in the Subgroup Population, as well as the IRT models, will be analyzed separately.

Table 3. 8. *Summary of study design.*

Conditions							
Population	**Test Type**	**Person Groups**	**IRT Model**	**Test Length**	**Priors**	**Final Trait Estimation**	**Theta Score Level**
Single Population	CT	All	Rasch	15	Population Composite Prior	EAP Normal	1 - 10
	CAT		2PL	30		EAP Uniform	
Subgroup Population	CT	A	Rasch	15	Population Composite Prior	EAP	1 - 10
		B_H	2PL	30	Group Composite Priors		
		B_L			Group Specific Priors		
					Less Informative Prior		
	CAT	A	Rasch	15	Population Composite Prior	EAP Normal	1 - 10
		B_H	2PL	30	Group Composite Priors	EAP Uniform	
		B_L			Group Specific Priors		

CHAPTER 4

RESULTS

This chapter presents the results of the simulation study. The chapter focuses on the estimates obtained from the simulation conditions, as well as the variables of interest discussed above (i.e., standard error, bias, and RMSE). The chapter begins first with a presentation of the simulation data and sample sets, then descriptive information for the conditions in each simulation. Next, correlation and regression results are presented, followed by the results of the split-plot ANOVAs conducted on the dependent measures. The major findings are presented in the ANOVA section for Simulation Two, Sample Set 2 data.

Before presentation of the results, a quick terminology review is warranted. Table 4.1 presents a quick reference of the language to be used in the remainder of this chapter. The table involves information relevant to both simulations, and then delves into information specific to the Simulation Two.

Table 4.1. *Terminology reference guide.*

Label	Definition	Explanation
	Pertinent to Both Simulations	
CT	Conventional Test	A static, fixed length test in which all items are administered in the same order to all simulees.
CAT	Computer Adaptive Test	An algorithmic testing design in which the test is specifically tailored to the simulee.
Test Length	15 or 30 items	Length of the test.
EAPN	EAP Normal	An ability estimate obtained using an informative prior in the estimation process.
Theta Score Level		Ten theta score levels, across the ability continuum, in which simulees are classified based on their true ability.
EAPU	EAP Uniform	An ability estimate obtained in which the prior utilized contains little information on the simulee.
	Simulation Two Specific	
A	Population group A (i.e., majority group)	This is the high-functioning population Group A described in Table 3.2, with a mean of 0.5.
B	Population group B (i.e., minority group)	This is the composite group B, composed of the low and high functioning subgroups (Table 3.2), with a mean of -0.3.
B_H	Group B High	This is the high ability subgroup of Group B (Table 3.2) with a mean of 0.5.
B_L	Group B Low	This is the low ability subgroup of Group B (Table 3.2) with a mean of -0.5.
Population Composite Prior	Prior that utilizes mean of the total population.	The mean of the population, computed using all groups/subgroups, is used as the test prior.
Group Composite Prior	Prior that utilizes mean of the groups from the population.	The mean of the two groups is computed and used from the subgroups. For A, there is only one subgroup. For Group B, this is the mean of B_H and B_L together.
Group Specific Prior	Prior that utilizes mean of the individual subgroups.	The mean of the subgroups is computed and used.
Less Informative Prior	Prior, only seen in CTs, where little prior information is used and bounds are placed.	The EAP Uniform estimate in CT designs.

Sample Sets

The simulation was ran using the design previously discussed – each theta score level was equally represented and each group, such as the three for Simulation Two, had the same number of simulees. This resulted in 3,000 simulees for a total of 300,000 across the 100 replications. However, to examine both the influence of theta score level and group membership in relation to the utilized prior information, three sets of data needed to be created. Sample Set 1 was a reduced version of the current data, in which all theta score levels were equally represented. This sample set was utilized in both simulation study analyses. Sample Set 2, used in both simulations, involved the peaked distributions of simulees, where the means were located at the same positions as in the population. For Simulation One, this distribution had a mean of 0 and SD of 1. Simulation Two involved consideration of the three groups with different true ability means. Thus, to look at the influence of the priors in a peaked distribution comparable to the population for Simulation Two, a sample was extracted from each condition to approximate the population distribution. Using weights obtained from Table 3.5, presented in Table 4.2, simulees were extracted from the total sample group; it is important to note that the means of the samples approximated those seen in the population. However, the group sizes were not proportional to that observed in the population for Sample Set 2. For example, in the total population, only 6% of the population is contained in B_L. To combat results being influenced by unequal group size, groups were constrained to have the same number of simulees. Then, for Simulation Two only, a third sample set (Sample Set 3) was obtained using weights in Table 4.3, in which not only the peaked distributions were obtained, but so was the total population group composition. Thus, Group B_L was 6% of the sample, Group B_H was 24%, and Group A

comprised the other 70%. This sample set was not analyzed for Simulation One because all groups were the same.

Table 4. 2. *Weights used for both simulations to approximate the peaked distributions.*

Theta	Simulation One		Simulation Two		
Score Level	A	B	A	B_L	B_H
1	0.0159	0.0071	0.0231	0.0658	0.0065
2	0.0310	0.0134	0.0323	0.0904	0.0155
3	0.0649	0.0273	0.0558	0.1523	0.0418
4	0.1049	0.0447	0.0733	0.1909	0.0913
5	0.1341	0.0571	0.0785	0.1913	0.1506
6	0.1343	0.0575	0.0684	0.1505	0.1958
7	0.1044	0.0451	0.0484	0.0927	0.1940
8	0.0641	0.0273	0.0278	0.0436	0.1494
9	0.0307	0.0134	0.0132	0.0163	0.0891
10	0.0157	0.0070	0.0078	0.0061	0.0660

Table 4. 3. *Weights used for Simulation Two to approximate the population distribution.*

Theta	Simulation Two		
Score Level	A	B_L	B_H
1	0.0043	0.0158	0.0004
2	0.0116	0.0217	0.0009
3	0.0308	0.0366	0.0025
4	0.0645	0.0458	0.0055
5	0.1049	0.0459	0.0090
6	0.1339	0.0361	0.0117
7	0.1339	0.0223	0.0116
8	0.1053	0.0105	0.0090
9	0.0636	0.0039	0.0053
10	0.0472	0.0015	0.0040

Descriptive Information

The following tables present the means and standard deviations for the total sample for a specific condition in both the datasets under each simulation. This information is presented for the true ability, EAP Normal, and EAP Uniform estimates. Both estimates are Bayesian; the difference lies in the fact that EAP Normal uses an informative prior, or a normal distribution

with a specific mean and standard deviation, whereas EAP Uniform uses a uniform distribution. For CTs, the EAP Uniform estimate is termed *Less Informative Prior* since it utilizes little information about the simulees and contains boundaries for the estimates. Table 4.4 presents the means and standard deviations for the ability estimates for all conditions under both IRT models for Simulation One. For Simulation Two, the means and standard deviations of combined subgroups for the three Sample Sets is presented, first for the CTs (Table 4.5) then the CATs (Table 4.6). However, for Sample Sets 2 and 3, the group means should approximate those seen in the population and warranted further investigation. Therefore, means for each condition by group (A, B_H, B_L) are presented in Tables 4.7 and 4.8 for the CT estimates for Sample Sets 2 and 3, respectively, and in Tables 4.9 through 4.12 for EAP Normal and EAP Uniform, respectively, for CAT for each sample sets.

For Simulation One, all means are approximately 0, as expected. For Simulation Two, Sample Sets 1 for both CT and CAT, the means of each condition is 0. This is as expected, since all theta score levels are equal. For Sample Sets 2 and 3 for each test type, however, the means are approximately 0.15 and 0.26, respectively. In Sample Set 2, the groups have the same number of simulees, but the distribution is like the population distribution (i.e., the mean of each group matches their true population mean). Thus, it is misleading to examine this table alone, and the means and SDs are given by groups. For Sample Set 3, the means and SDs are also given by groups to ensure that the population was represented correctly. In these tables (4.7 – 4.12), the ability estimates for a group are close to the true mean. An interesting find is that, for Group B_L using a CT (Tables 4.7 and 4.8), the estimates are overestimated. This is because the mean of the CT is much higher than the mean of the group. Thus, there were not enough items to accurately estimate this group's ability. Tables 4.9 through 4.12 provide the most vital

information, as these tables present the information for the individual test priors by subgroups for EAP Normal and EAP Uniform estimates, respectively. The means are approximately similar for each of the subgroups when compared to the true ability. The lowest means are obtained under the Rasch model for Group B_H when using the Group Composite Prior under both final ability estimate approaches. When using this prior, its informative mean (M = -0.3) is far away from Group B_H's mean (M = 0.5). Therefore, this prior, when utilized for test administration and final ability estimation, has more of an influence on the final ability estimates. For the Population Composite Prior, the lowest means are obtained for Group BL under the Rasch model. The mean of this informative prior (M = 0.26) is much higher than the mean ability of Group BL (M = -0.5). The 2PL model can recover ability better due to the higher discrimination values, which lead to more informative item selection.

Table 4. 4. *Descriptive statistics for Simulation One for both sample sets for each IRT model, test prior, and test length (*N *= 9,900).*

IRT Model	Test Type	Test Length	Sample Set 1 – Uniform $\overline{\theta_{True}} = -0.025;\ SD_{\theta_{True}} = 1.460$				Sample Set 2 – Peaked $\overline{\theta_{True}} = 0.008;\ SD_{\theta_{True}} = 1.033$			
			$\widehat{\theta_{EAP\ Normal}}$		$\widehat{\theta_{EAP\ Uniform}}$		$\widehat{\theta_{EAP\ Normal}}$		$\widehat{\theta_{EAP\ Uniform}}$	
			M	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
Rasch	CT	15	-0.020	1.072	-0.033	1.389	0.016	0.831	0.016	1.044
		30	-0.027	1.081	-0.024	1.237	-0.013	0.809	-0.010	0.908
	CAT	15	-0.027	1.181	-0.031	1.343	0.007	0.895	0.007	1.008
		30	-0.020	1.293	-0.021	1.370	0.016	0.946	0.017	0.999
2PL	CT	15	-0.033	1.230	-0.050	1.507	-0.001	0.917	-0.003	1.084
		30	-0.001	1.249	-0.009	1.405	0.007	0.921	0.008	1.010
	CAT	15	-0.023	1.369	-0.024	1.451	0.008	1.008	0.009	1.065
		30	-0.025	1.391	-0.026	1.446	0.005	1.016	0.006	1.055

Table 4. 5. *Descriptive statistics for combined subgroups in Simulation Two for estimated ability for both sample sets for the conventional tests, or CTs, for each IRT model, test prior, and test length.*

			Sample Set 1 – Uniform Distribution $\overline{\theta_{True}} = 0.004$ $SD_{\theta_{True}} = 1.474$ $N = 30{,}000$		**Sample Set 2 – Peaked Distribution** $\overline{\theta_{True}} = 0.164$ $SD_{\theta_{True}} = 1.122$ $N = 30{,}000$		**Sample Set 3 – Population Distribution** $\overline{\theta_{True}} = 0.259$ $SD_{\theta_{True}} = 1.083$ $N = 9{,}900$	
IRT Model	**Final Estimate Prior**	**Test Length**	$\widehat{\theta}$		$\widehat{\theta}$		$\widehat{\theta}$	
			M	***SD***	***M***	***SD***	***M***	***SD***
Rasch Model	Population Composite	15	0.130	0.941	0.207	0.767	0.263	0.738
		30	0.103	1.010	0.205	0.780	0.272	0.780
	Group Composite	15	0.067	0.957	0.170	0.785	0.263	0.768
		30	0.084	1.021	0.191	0.808	0.275	0.792
	Group Specific	15	0.097	0.957	0.193	0.805	0.263	0.778
		30	0.101	1.020	0.203	0.820	0.274	0.797
	Less Informative	15	0.037	1.313	0.181	1.037	0.260	0.990
		30	0.074	1.192	0.199	0.920	0.274	0.891
2PL Model	Population Composite	15	0.089	1.093	0.190	0.881	0.255	0.855
		30	0.069	1.157	0.186	0.915	0.268	0.888
	Group Composite	15	0.031	1.114	0.158	0.902	0.253	0.889
		30	0.055	1.169	0.180	0.926	0.270	0.904
	Group Specific	15	0.055	1.113	0.172	0.921	0.251	0.899
		30	0.067	1.169	0.185	0.938	0.268	0.910
	Less Informative	15	-0.049	1.608	0.140	1.262	0.241	1.224
		30	0.013	1.490	0.171	1.148	0.270	1.101

Table 4. 6. *Descriptive statistics for combined subgroups in Simulation Two for estimated ability, under both final estimation approaches, for both sample sets for the computer adaptive tests, or CATs, for each IRT model, test prior, and test length.*

			Sample Set 1 – Uniform $\overline{\theta_{True}} = 0.006$ $SD_{\theta_{True}} = 1.477$ N = 30,000				**Sample Set 2 – Peaked** $\overline{\theta_{True}} = 0.165$ $SD_{\theta_{True}} = 1.123$ N = 30,000				**Sample Set 3 – Population** $\overline{\theta_{True}} = 0.259$ $SD_{\theta_{True}} = 1.082$ N = 9,900			
IRT Model	**Test Prior**	**Test Length**	$\widehat{\theta_{EAP\ Normal}}$		$\widehat{\theta_{EAP\ Uniform}}$		$\widehat{\theta_{EAP\ Normal}}$		$\widehat{\theta_{EAP\ Normal}}$		$\widehat{\theta_{EAP\ Normal}}$		$\widehat{\theta_{EAP\ Normal}}$	
			M	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***
Rasch Model	Population Composite	15	0.072	1.188	0.044	1.352	0.186	0.954	0.176	1.078	0.258	0.931	0.258	1.050
		30	0.053	1.300	0.039	1.377	0.179	1.015	0.174	1.073	0.261	0.984	0.260	1.039
	Group Composite	15	-0.001	1.194	0.002	1.356	0.123	0.973	0.142	1.086	0.252	0.966	0.252	1.068
		30	0.004	1.33	0.006	1.380	0.140	1.026	0.150	1.079	0.254	1.002	0.254	1.050
	Group Specific	15	0.047	1.186	0.028	1.345	0.167	1.003	0.167	1.103	0.258	0.974	0.259	1.070
		30	0.031	1.305	0.022	1.382	0.168	1.049	0.168	1.096	0.254	1.010	0.255	1.054
2PL Model	Population Composite	15	0.023	1.392	0.010	1.477	0.172	1.084	0.168	1.146	0.262	1.046	0.263	1.106
		30	0.017	1.410	0.008	1.467	0.169	1.089	0.166	1.132	0.253	1.049	0.253	1.090
	Group Composite	15	0.003	1.389	0.005	1.473	0.150	1.087	0.160	1.145	0.254	1.062	0.255	1.116
		30	-0.002	1.403	-0.001	1.459	0.152	1.093	0.160	1.132	0.250	1.056	0.251	1.091
	Group Specific	15	0.020	1.395	0.011	1.478	0.167	1.096	0.168	1.148	0.259	1.060	0.261	1.110
		30	0.009	1.402	0.003	1.458	0.162	1.097	0.162	1.133	0.255	1.064	0.256	1.097

Table 4. 7. *Descriptive statistics, by group, for Simulation Two for the CT Sample Set 2 ability estimates for each IRT model, final estimate test prior, and test length.*

IRT Model	Final Estimate Prior	Test Length	A $\overline{\theta_{True}} = 0.492$ $SD_{\theta_{True}} = 1.016$ N = 10,000		B_H $\overline{\theta_{True}} = 0.496$ $SD_{\theta_{True}} = 1.026$ N = 10,000		B_L $\overline{\theta_{True}} = -0.495$ $SD_{\theta_{True}} = 1.020$ N = 10,000	
			M	***SD***	***M***	***SD***	***M***	***SD***
Rasch Model	Population Composite	15	0.410	0.715	0.404	0.732	-0.192	0.689
		30	0.432	0.734	0.410	0.728	-0.226	0.755
	Group Composite	15	0.437	0.719	0.340	0.701	-0.266	0.742
		30	0.445	0.736	0.392	0.731	-0.265	0.755
	Group Specific	15	0.437	0.719	0.433	0.717	-0.290	0.751
		30	0.445	0.736	0.442	0.742	-0.279	0.760
	Less Informative	15	0.450	0.932	0.447	0.931	-0.352	1.034
		30	0.446	0.832	0.444	0.842	-0.294	0.879
2PL Model	Population Composite	15	0.430	0.817	0.443	0.823	-0.301	0.787
		30	0.456	0.839	0.440	0.836	-0.337	0.834
	Group Composite	15	0.450	0.822	0.388	0.810	-0.363	0.837
		30	0.465	0.841	0.438	0.833	-0.362	0.854
	Group Specific	15	0.450	0.822	0.450	0.819	-0.385	0.854
		30	0.465	0.841	0.466	0.843	-0.377	0.866
	Less Informative	15	0.493	1.117	0.492	1.105	-0.564	1.252
		30	0.495	1.019	0.497	1.018	-0.478	1.116

Table 4. 8. *Descriptive statistics, by group, for Simulation Two for the CT Sample Set 3 ability estimates for each IRT model, final estimate test prior, and test length.*

IRT Model	Final Estimate Prior	Test Length	A $\overline{\theta_{True}} = 0.550$ $SD_{\theta_{True}} = 0.983$ N = 6,800		B_H $\overline{\theta_{True}} = 0.497$ $SD_{\theta_{True}} = 0.854$ N = 600		B_L $\overline{\theta_{True}} = -0.589$ $SD_{\theta_{True}} = 0.939$ N = 2,500	
			M	***SD***	***M***	***SD***	***M***	***SD***
Rasch Model	Population Composite	15	0.441	0.686	0.384	0.686	-0.249	0.644
		30	0.469	0.712	0.390	0.656	-0.291	0.708
	Group Composite	15	0.467	0.690	0.376	0.633	-0.317	0.701
		30	0.483	0.715	0.418	0.628	-0.325	0.718
	Group Specific	15	0.467	0.690	0.465	0.659	-0.341	0.712
		30	0.483	0.715	0.466	0.644	-0.339	0.723
	Less Informative	15	0.490	0.886	0.486	0.809	-0.420	0.987
		30	0.492	0.806	0.467	0.699	-0.363	0.845
2PL Model	Population Composite	15	0.471	0.793	0.430	0.740	-0.373	0.724
		30	0.500	0.812	0.443	0.745	-0.408	0.762
	Group Composite	15	0.492	0.799	0.418	0.754	-0.436	0.786
		30	0.511	0.817	0.456	0.742	-0.429	0.795
	Group Specific	15	0.492	0.799	0.479	0.771	-0.458	0.805
		30	0.511	0.817	0.480	0.752	-0.443	0.808
	Less Informative	15	0.550	1.074	0.534	0.985	-0.669	1.206
		30	0.554	0.979	0.500	0.826	-0.555	1.055

Table 4. 9. *Descriptive statistics, by group, for Simulation Two for the CAT Sample Set 2 ability estimates for each IRT model, test prior, and test length for EAP Normal ability estimates.*

IRT Model	Test Prior	Test Length	A $\overline{\theta_{True}} = 0.497$ $SD_{\theta_{True}} = 1.022$ N = 10,000		BH $\overline{\theta_{True}} = 0.494$ $SD_{\theta_{True}} = 1.021$ N = 10,000		BL $\overline{\theta_{True}} = -0.495$ $SD_{\theta_{True}} = 1.021$ N = 10,000	
			M	***SD***	***M***	***SD***	***M***	***SD***
Rasch Model	Population Composite	15	0.439	0.889	0.439	0.884	-0.319	0.882
		30	0.459	0.935	0.464	0.934	-0.386	0.930
	Group Composite	15	0.492	0.887	0.313	0.873	-0.437	0.897
		30	0.497	0.940	0.383	0.925	-0.460	0.933
	Group Specific	15	0.493	0.888	0.493	0.892	-0.485	0.893
		30	0.503	0.936	0.494	0.938	-0.494	0.943
2PL Model	Population Composite	15	0.475	1.001	0.474	0.995	-0.434	0.991
		30	0.481	0.995	0.477	0.998	-0.453	0.999
	Group Composite	15	0.492	0.998	0.434	0.983	-0.476	0.996
		30	0.488	1.001	0.450	0.988	-0.482	1.000
	Group Specific	15	0.504	0.994	0.489	0.992	-0.493	0.989
		30	0.488	0.997	0.486	0.994	-0.489	0.998

Table 4. 10. Descriptive statistics, by group, for Simulation Two for the CAT Sample Set 2 ability estimates for each IRT model, test prior, and test length for EAP Uniform ability estimates.

IRT Model	Test Prior	Test Length	A $\overline{\theta_{True}} = 0.497$ $SD_{\theta_{True}} = 1.022$ N = 10,000		BH $\overline{\theta_{True}} = 0.494$ $SD_{\theta_{True}} = 1.021$ N = 10,000		BL $\overline{\theta_{True}} = -0.495$ $SD_{\theta_{True}} = 1.021$ N = 10,000	
			M	***SD***	***M***	***SD***	***M***	***SD***
Rasch Model	Population Composite	15	0.461	1.002	0.461	0.995	-0.393	1.002
		30	0.470	0.988	0.475	0.986	-0.422	0.985
	Group Composite	15	0.492	0.998	0.389	0.991	-0.455	1.011
		30	0.497	0.993	0.422	0.980	-0.469	0.985
	Group Specific	15	0.493	0.999	0.492	1.003	-0.485	1.004
		30	0.504	0.988	0.494	0.990	-0.495	0.995
2PL Model	Population Composite	15	0.488	1.058	0.487	1.052	-0.472	1.050
		30	0.491	1.033	0.486	1.037	-0.479	1.038
	Group Composite	15	0.493	1.055	0.475	1.041	-0.487	1.055
		30	0.489	1.039	0.480	1.027	-0.488	1.039
	Group Specific	15	0.506	1.050	0.491	1.049	-0.494	1.047
		30	0.489	1.035	0.487	1.032	-0.489	1.037

Table 4. 11. *Descriptive statistics, by group, for Simulation Two for the CAT Sample Set 3 ability estimates for each IRT model, test prior, and test length for EAP Normal ability estimates.*

IRT Model	Test Prior	Test Length	A $\overline{\theta_{True}} = 0.549$ $SD_{\theta_{True}} = 1.982$ N = 6,800		B_H $\overline{\theta_{True}} = 0.500$ $SD_{\theta_{True}} = 1.845$ N = 600		B_L $\overline{\theta_{True}} = -0.588$ $SD_{\theta_{True}} = 1.940$ N = 2,500	
			M	***SD***	***M***	***SD***	***M***	***SD***
Rasch Model	Population Composite	15	0.482	0.866	0.442	0.795	-0.396	0.819
		30	0.505	0.909	0.483	0.783	-0.457	0.865
	Group Composite	15	0.528	0.868	0.309	0.758	-0.512	0.845
		30	0.534	0.907	0.401	0.782	-0.543	0.862
	Group Specific	15	0.533	0.862	0.528	0.771	-0.554	0.844
		30	0.540	0.900	0.488	0.807	-0.578	0.867
2PL Model	Population Composite	15	0.528	0.957	0.479	0.854	-0.513	0.930
		30	0.527	0.956	0.489	0.850	-0.549	0.915
	Group Composite	15	0.541	0.960	0.430	0.859	-0.570	0.936
		30	0.531	0.961	0.459	0.853	-0.564	0.918
	Group Specific	15	0.544	0.956	0.508	0.871	-0.573	0.922
		30	0.541	0.962	0.491	0.856	-0.581	0.926

Table 4. 12. *Descriptive statistics, by group, for Simulation Two for the CAT Sample Set 3 ability estimates for each IRT model, test prior, and test length for EAP Uniform ability estimates.*

IRT Model	Test Prior	Test Length	A $\overline{\theta_{True}} = 0.549$ $SD_{\theta_{True}} = 1.982$ $N = 6{,}800$		B_H $\overline{\theta_{True}} = 0.500$ $SD_{\theta_{True}} = 1.845$ $N = 600$		B_L $\overline{\theta_{True}} = -0.588$ $SD_{\theta_{True}} = 1.940$ $N = 2{,}500$	
			M	***SD***	***M***	***SD***	***M***	***SD***
Rasch Model	Population Composite	15	0.510	0.974	0.464	0.888	-0.480	0.932
		30	0.519	0.960	0.494	0.823	-0.498	0.917
	Group Composite	15	0.533	0.976	0.381	0.853	-0.542	0.952
		30	0.536	0.957	0.439	0.825	-0.558	0.910
	Group Specific	15	0.538	0.968	0.531	0.857	-0.565	0.947
		30	0.543	0.949	0.487	0.849	-0.584	0.914
2PL Model	Population Composite	15	0.544	1.011	0.492	0.897	-0.556	0.987
		30	0.538	0.993	0.499	0.881	-0.579	0.952
	Group Composite	15	0.545	1.014	0.469	0.905	-0.587	0.993
		30	0.533	0.997	0.488	0.885	-0.574	0.953
	Group Specific	15	0.548	1.010	0.509	0.916	-0.579	0.977
		30	0.544	0.998	0.491	0.886	-0.584	0.962

Correlations

Correlations were obtained for every condition between the true score and EAP Normal estimate, as well as between the true score and EAP Uniform estimate. Regressions were conducted for each replication and the R^2 extracted from each analysis. The mean was taken across all R^2 values, and then was square rooted. Table 4.9 shows the correlations for Simulation One; Table 4.10 shows the correlations for Simulation Two CT; and Table 4.11 shows the correlations for Simulation Two CAT. In general, correlations are higher for estimates obtained using a CAT than a CT. Also, they are higher when obtained using an informative prior (i.e., EAP Normal) than a less informative prior (i.e., EAP Uniform). But all these differences are negligible. Lastly, correlations are higher for estimates obtained using the 2PL model since the Rasch model has lower discrimination values.

Table 4. 13. *Correlations between true ability and estimated abilities for all conditions in both sample sets for Simulation One.*

				Data Set 1 - Uniform		**Data Set 2 - Peaked**	
Prior	**Test Type**	**Test Length**	**IRT Model**	r_{Norm}	r_{Uni}	r_{Norm}	r_{Uni}
Population Composite Prior	CT	15	Rasch	0.918	0.910	0.852	0.852
			2PL	0.958	0.948	0.926	0.922
		30	Rasch	0.954	0.950	0.919	0.918
			2PL	0.979	0.973	0.964	0.962
	CAT	15	Rasch	0.936	0.934	0.881	0.882
			2PL	0.973	0.973	0.950	0.950
		30	Rasch	0.966	0.966	0.937	0.937
			2PL	0.982	0.982	0.965	0.965

Table 4. 14. *Correlations between true ability and estimated ability for all conditions involving final estimate prior, test length, and IRT model in all three samples sets for Simulation Two, CT.*

Final Estimate Prior	Test Length	IRT Model	$r_{SampleSet1}$	$r_{SampleSet2}$	$r_{SampleSet3}$
Population Composite Prior	15	Rasch	0.924	0.884	0.876
		2PL	0.956	0.939	0.926
	30	Rasch	0.955	0.929	0.884
		2PL	0.974	0.968	0.928
Group Composite Prior	15	Rasch	0.925	0.890	0.885
		2PL	0.957	0.940	0.928
	30	Rasch	0.956	0.930	0.868
		2PL	0.975	0.968	0.922
Group Specific Prior	15	Rasch	0.925	0.894	0.938
		2PL	0.957	0.942	0.968
	30	Rasch	0.956	0.931	0.940
		2PL	0.975	0.969	0.969
Less Informative Prior	15	Rasch	0.909	0.878	0.940
		2PL	0.936	0.917	0.969
	30	Rasch	0.948	0.925	0.914
		2PL	0.956	0.948	0.950

Table 4. 15. *Correlations between true ability and estimate ability for all conditions involving test prior, test length, and IRT model in both sample sets for Simulation Two, CAT.*

Test Prior	Test Length	IRT Model	Data Set 1 - Uniform		Data Set 2 - Peaked		Data Set 3 - Population	
			r_{Norm}	r_{Uni}	r_{Norm}	r_{Uni}	r_{Norm}	r_{Uni}
Population Composite Prior	15	Rasch	0.937	0.935	0.895	0.895	0.890	0.891
		2PL	0.974	0.974	0.956	0.957	0.942	0.942
	30	Rasch	0.968	0.968	0.945	0.945	0.894	0.893
		2PL	0.983	0.983	0.970	0.970	0.944	0.943
Group Composite Prior	15	Rasch	0.932	0.933	0.897	0.898	0.899	0.898
		2PL	0.974	0.974	0.956	0.956	0.943	0.943
	30	Rasch	0.966	0.967	0.944	0.945	0.954	0.954
		2PL	0.983	0.983	0.969	0.969	0.967	0.968
Group Specific Prior	15	Rasch	0.932	0.933	0.901	0.900	0.954	0.954
		2PL	0.974	0.974	0.957	0.956	0.967	0.967
	30	Rasch	0.966	0.966	0.947	0.947	0.954	0.953
		2PL	0.982	0.982	0.970	0.970	0.968	0.968

Regression Analysis

Regressions were conducted using ten randomly selected replications from each condition to graphically display information. Replications 4, 11, 26, 27, 34, 35, 39, 48, 76, and 90 was randomly selected from all 100 replications for use in the regression analyses. The 30-item test under the 2PL model was used to conduct all regression analyses. For Simulation One, Sample Sets 1 and 2 are presented in Figures 4.1 and Figure 4.2, respectively. For Simulation Two, the Group Composite Prior was used for the regressions. If the approach is used in practice, this is the prior that would be utilized. The Group Specific Prior cannot be applied, since there is no way to know an individual's true ability to accurate place them in a high or low functioning subgroup. For CTs, the regressions are displayed in Figure 4.3 using the final estimate prior. Figure 4.4 shows the results for all sample sets data for the CAT.

Figure 4.1 presents the regression analyses for Simulation One using Sample Set 1, where all theta score levels are equally represented. As shown, the true abilities and estimated abilities are relatively similar. There is a slight curvi-linear relationship towards the ability continuum extremes for the CTs (A and B); simulees at these extremes have a high chance for ability-difficulty mismatch when using a CT, thus resulting in a higher degree of inaccuracy when estimating ability. However, this pattern disappears when using a CAT, since the test is tailored to the simulee.

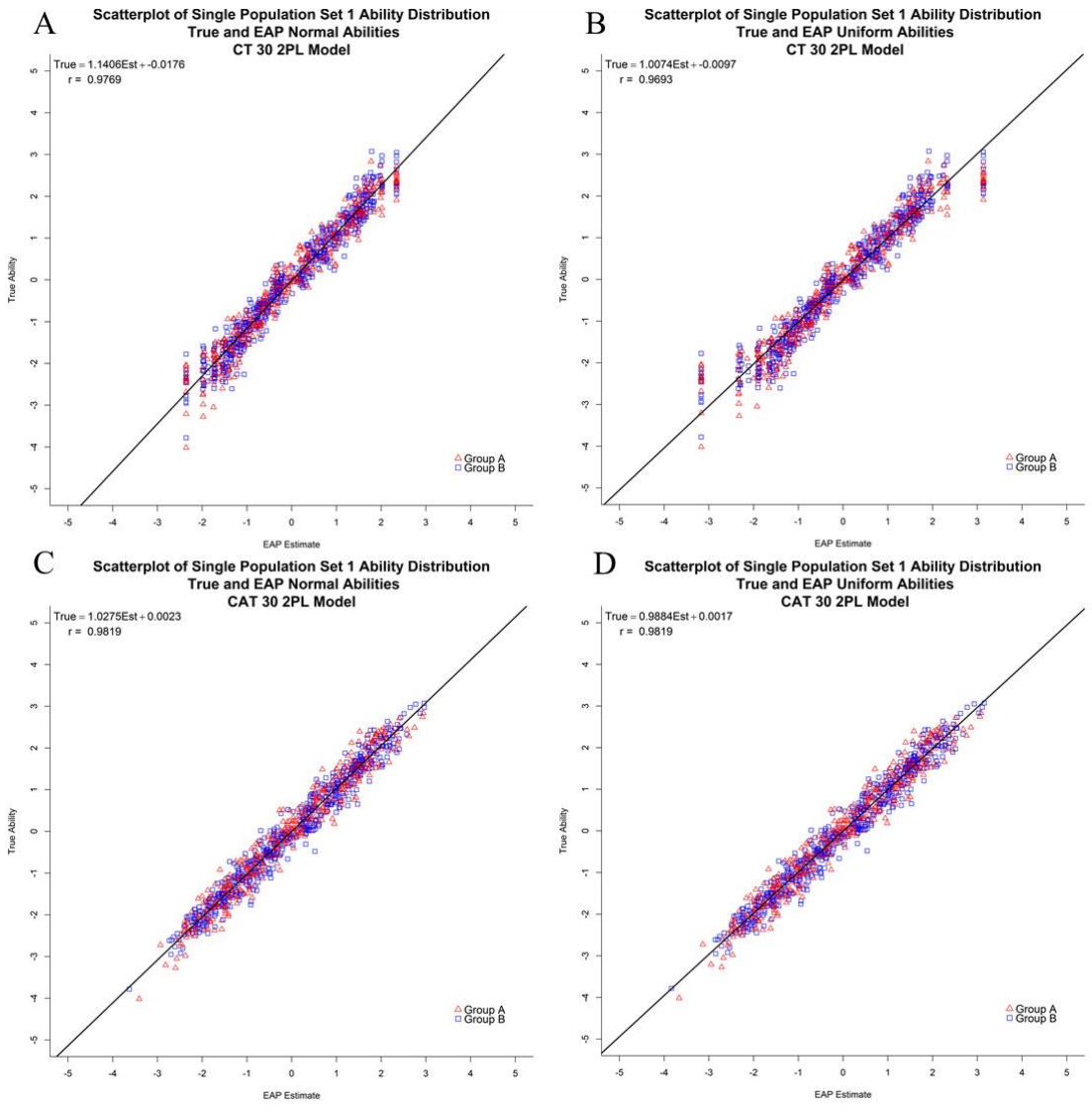

Figure 4. 1. Regression plots for EAP Normal and EAP Uniform estimates for the Single Population under the 2PL model for the 30-item CT and CAT, where Group A is represented in red and Group B is blue, for Sample Set 1.

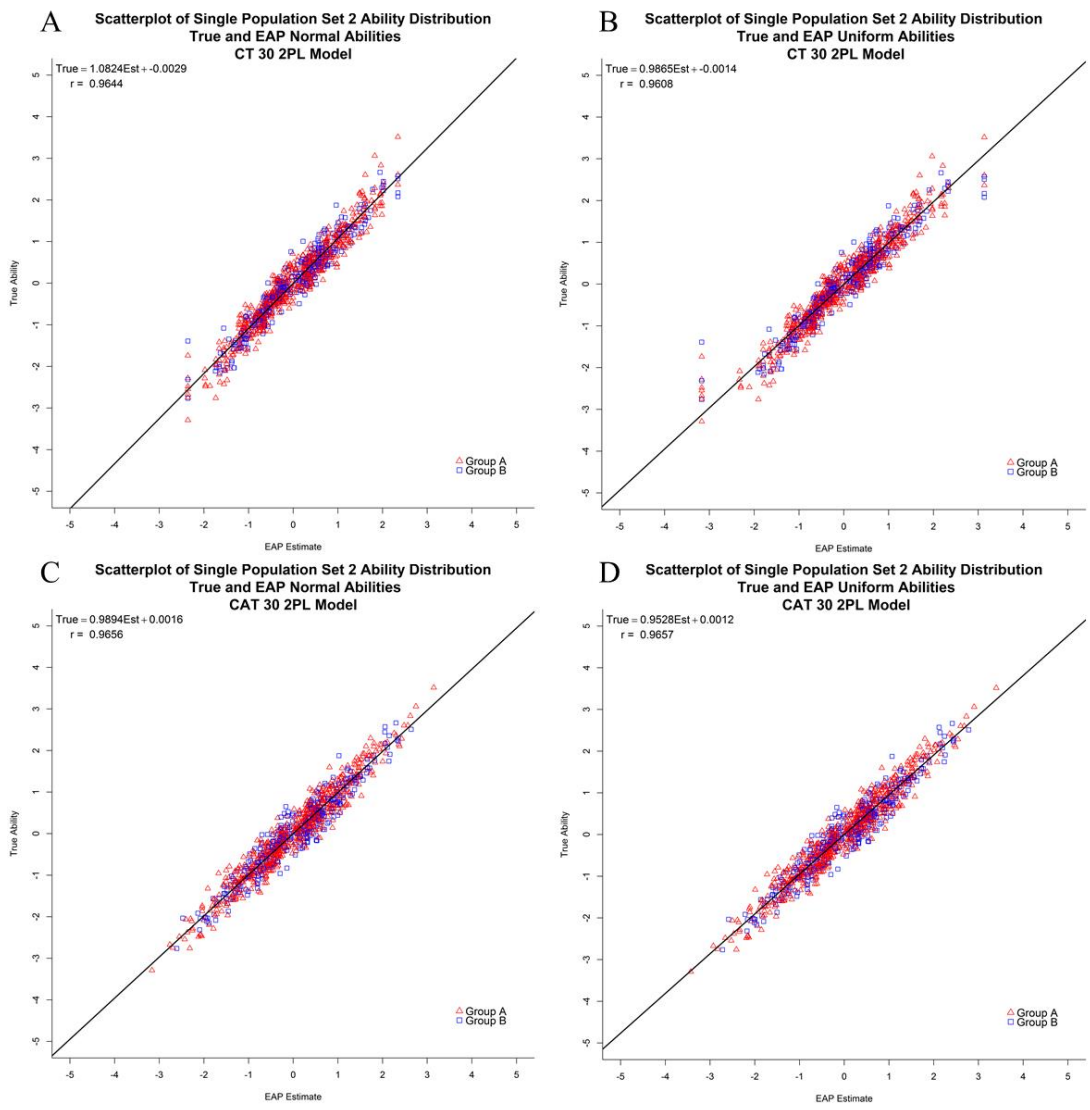

Figure 4. 2. Regression plots for EAP Normal and EAP Uniform estimates for the Single Population under the 2PL model for the 30-item CT and CAT, where Group A is represented in red and Group B is blue, for Sample Set 2.

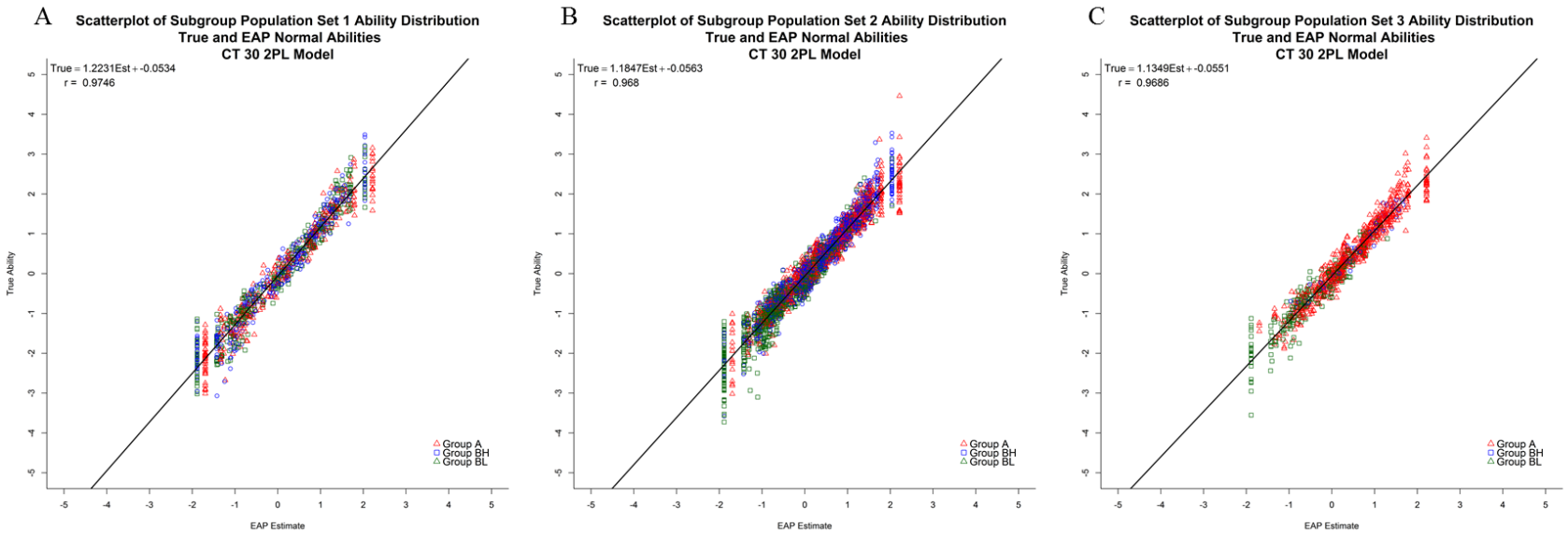

Figure 4. 3. Regression plots for EAP estimates using the Group Composite Prior for final ability estimation for the Subgroup Population under the 2PL model for the 30-item CT, where Group A is represented by red triangles, Group BH is blue circles, and Group BL is green squares, for Sample Sets 1, 2, and 3.

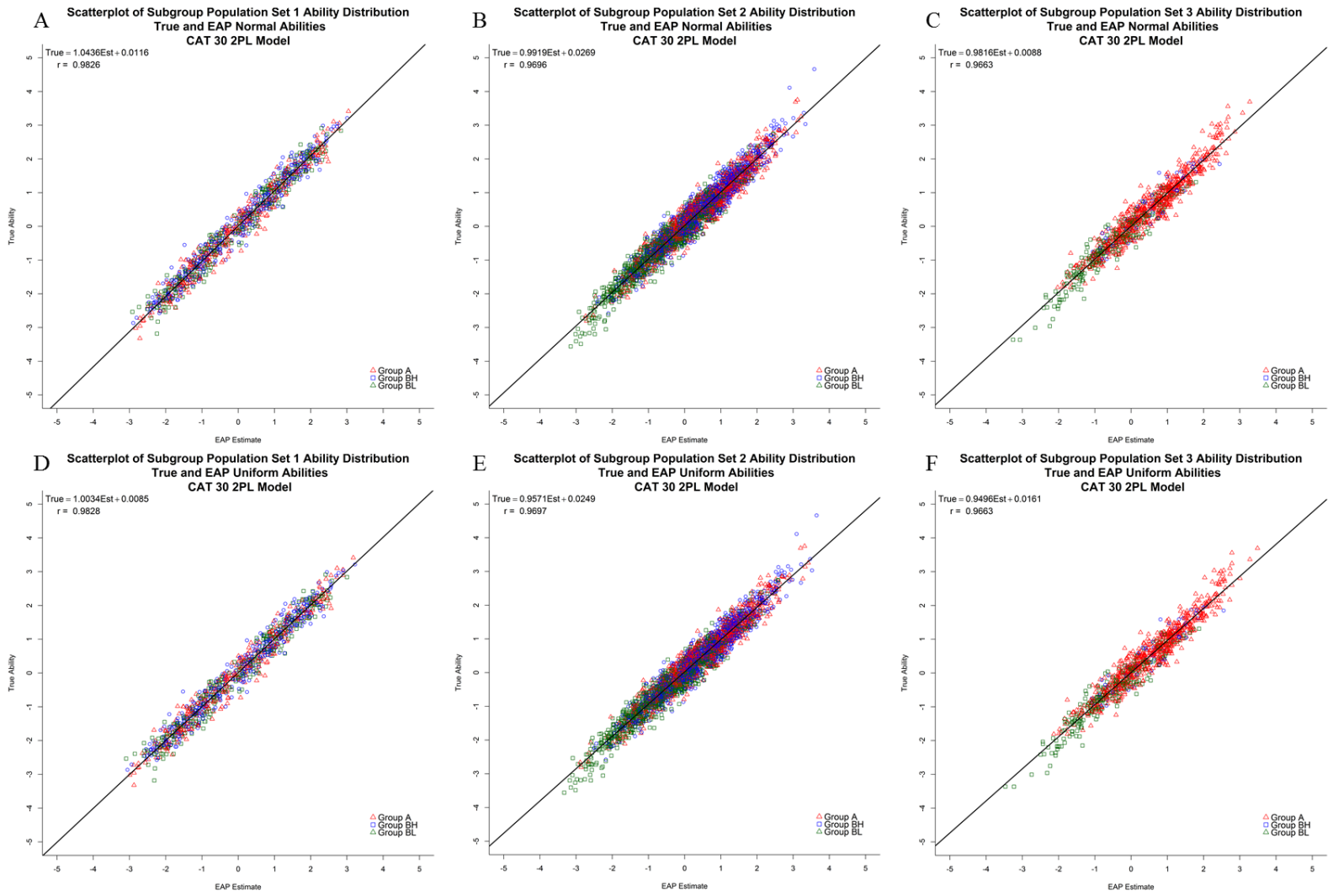

Figure 4. 4. Regression plots for EAP Normal and EAP Uniform estimates for the Subgroup Population under the 2PL model for the 30-item CAT, where Group A is represented by red triangles, Group BH is blue circles, and Group BL is green squares, for Sample Sets 1, 2, and 3.

ANOVA Analyses

Multiple split plot ANOVAs were conducted. For Simulation One, two sets of ANOVAs were conducted. For Simulation Two, three sets of ANOVAs were conducted; the first two are similar to those in Simulation One, and the last used Sample Set 3, which was utilized only in the second simulation. The first set of ANOVAs (i.e., ANOVA Set 1) used the Sample Set 1 data, in which all theta score levels are equally represented. The second set of ANOVAs (i.e., ANOVA Set 2) used the Sample Set 2 sample data, in which the distributions for the subgroups reflect that observed in the population. The third set of ANOVAs for Simulation Two (i.e., ANOVA Set 3) used the Sample Set 3 sample data, in which the sample reflected the true make-up of the population. This set of ANOVAs examines group membership. Data were examined these ways because equally represented theta score groups may confound the effects that might be observed when using the individual test priors. When examining results at the theta score level, the mean being used is that of the score level of interest, not the group. Therefore, the first set of ANOVAs for each simulation reflected how the theta score levels and test prior interact. Thus, the second and third set of ANOVAs will highlight the interaction between the test prior and group membership using Sample Sets 2 and 3. For each set of ANOVAs, IRT models (R = Rasch model; 2pl = 2PL model) were examined separately and not as another factor. Therefore, the data were split on this variable for analyses. While the Rasch model is presented in conjunction with the 2PL model for both simulations, it is pertinent to note that the two IRT models should not be compared to each other. For the comparison to be appropriate, the single Rasch model item discrimination parameter should be estimated to be a constant instead of constrained to 1. Therefore, comparisons between the two models were not made.

For Simulation One, or the Single Population, ANOVA Set 1 involved six Test Type (2) x Test Length (2) x Trait Estimate Type (2) x Theta Score Level (10) split-plot designs, where the first two factors are the between replication factors and the last two are the within replication factors. For these ANOVAs, replications are treated as subjects as in a traditional split-plot design. The between replication factors are test type (T = CAT or CT) and test length (L = 15 or 30). The within replication factors are final trait estimate type (E = EAP Normal or EAP Uniform) and theta score level (S = 1 to 10). The Sample Set 1 data were used, in which all theta score levels are equally represented. ANOVA Set 2 examined a Test Type (2) x Test Length (2) x Trait Estimate Type (2) design. This second set utilized the Sample Set 2 data, in which the population distribution is observed. A cross between test type and estimate type is appropriate for Simulation One because there are only two estimate types, one which uses an informative prior (EAP Normal) and one which uses a less informative one (EAP Uniform).

A different approach had to be taken for Simulation Two, which examined the Subgroup Population. First, IRT models are examined separately, as previously mentioned. However, test type had to be examined individually as well. For a CAT, three test priors existed, whereas for the CTs, four test priors existed. For the CAT, the test priors are used as a first estimate of ability and to select items. Then, at the end, final trait estimation is conducted using EAP Normal, which uses the test prior during the test, or EAP Uniform, which uses a less informative prior. Therefore, they are called test priors. For the CTs, either the informative prior was used at the end of the test to obtain an estimate or a less informative prior was used. It is not appropriate to have a cross of estimate type and test prior for the CTs since the test prior is only utilized at the end. The priors only influence final ability estimation and are thus referred to as final estimate priors. Thus, each test type had both to be examined separately.

For CT ANOVA Set 1, six Test Length (2) x Final Estimate Prior (4) x Theta Score Level (10) split-plot ANOVAs were conducted, where the first two factors are the between subject factors and the remaining one is a within replication factor. Between replication factors for this set are test length (L = 15 or 30) and final estimate prior (P = Population Composite Prior, Group Composite Prior, Group Specific Prior, Less Informative Prior). The within replication factor is theta score level (S = 1 to 10). For CT ANOVA Set 2, six Test Length (2) x Final Estimate Prior (4) x Group Membership (3) split-plot ANOVAs are conducted. Here, the first two factors are the same as in Set 1. However, the last factor is the only within replication factor and is group membership (G = A, B_H, B_L). Theta score level and group membership are examined separately to ensure effects are not masked. For CT ANOVA Set 3, the same six split-plot ANOVAs as in Set 2 are conducted, except Sample Set 3 is utilized.

For the CAT analyses, both sets of data were examined. For CAT ANOVA Set 1, six Test Length (2) x Test Prior (3) x Trait Estimate Type (2) x Theta Score Level (10) split-plot ANOVAs were conducted, where the first two factors are the between subject factors and the remaining two are within subject factors. Between replication factors for this set are test length (L = 15 or 30) and test prior (P = Population Composite Prior, Group Composite Prior, Group Specific Prior). Within replication factors are final trait estimate type (E = EAP Normal or EAP Uniform) and theta score level (S = 1 to 10). Lastly, for CAT ANOVA Set 2 for this simulation, a Test Length (2) x Test Prior (3) x Trait Estimate Type (2) x Group Membership (3) design was employed. The same two between replication factors are utilized. However, the two within replication factors are estimate type (E = EAP Normal and EAP Uniform) and group membership (G = A, B_H, $_{BL}$). For CAT ANOVA Set 3, as with the CTs, the design from ANOVA Set 2 is used but with Sample Set 3 data.

The following dependent measures were examined within each ANOVA set: mean standard error (SE), bias, and root mean square error (RMSE). Since there are three DVs, the family-wise error rate was controlled; therefore, $p = \frac{0.05}{3} \approx 0.0167$. Violations of sphericity are corrected using a Huynh-Feldt adjustment (Huynh & Feldt, 1976). Lastly, within-family η^2 (i.e., η_w^2; Roberts & Thompson, 2011) was calculated for each effect using effect families in the study design. For example, one effect family is the between-factor error term; another effect family is within-factor error term related to those tested by the final trait estimate type. Thus, for the Subgroup Population, eight effect families existed. Only those effects which are statistically significant ($p < 0.0167$) and had a η_w^2 value greater than 0.07 are interpreted.

Results for the ANOVAs are organized as follows. First, Simulation One (i.e., Single Population) results are presented, serving as a base rate comparison for Simulation Two (i.e., Subgroup Population), which contains the subgroups of interest. Within this section, ANOVA Set 1 is presented first, followed by ANOVA Set 2. After presentation of these results, the results of Simulation Two are presented. As with the first simulation, the order of presentation is ANOVA Set 1, ANOVA Set 2, then ANOVA Set 3. However, these sets are presented within their respective test types, where CT is presented first, followed by CAT. With this simulation, the influence of prior is of primary interest, and thus these results are presented first for each group. Then, all other noteworthy results are presented.

Simulation One

This section utilizes the Single Population data, in which there are no groups. The distribution for the population is a standard normal distribution ($\theta \sim N(0, 1)$).

ANOVA Set 1. This first set of ANOVAs utilized the sample Set 1, or uniform sample. In this set, all theta score levels are equally represented to examine SE, bias, and RMSE at

different slices of the ability continuum. This was done because examination of the theta score levels in conjunction with a peaked distribution will confound effects. The η_w^2 for all effects and interactions for each dependent variable, separated by IRT model, is presented in Table 4.16.

Table 4. 16. *Within-family effect sizes from ANOVA Set 1 conducted on standard error (SE), bias, root mean square error (RMSE) for each IRT model using the Single Population.*

Effect	SE		Bias		RMSE	
	Rasch	2PL	Rasch	2PL	Rasch	2PL
T	**0.392****	**0.460****	0.001	0.000	**0.294****	**0.370****
L	**0.599****	**0.438****	0.004	0.065**	**0.573****	**0.468****
T*L	0.007**	**0.094****	0.002	**0.079****	0.005**	0.049**
E	**0.712****	**0.613****	0.010	**0.101****	**0.573****	0.043**
E*T	**0.130****	**0.324****	0.001	**0.104****	0.043**	**0.084****
E*L	**0.143****	0.039**	**0.109****	0.029**	0.037**	**0.161****
E*T*L	0.011**	0.018**	**0.079****	0.025**	**0.089****	**0.129****
S	**0.674****	**0.512****	**0.795****	**0.558****	**0.482****	**0.342****
S*T	**0.169****	**0.389****	0.037**	0.068**	**0.072****	**0.206****
S*L	0.023**	0.009**	0.005**	0.002*	0.004**	0.013**
S*T*L	0.016**	0.011**	0.011**	0.006**	0.016**	0.014**
E*S	**0.635****	**0.506****	**0.797****	**0.704****	**0.605****	**0.178****
E*S*T	**0.175****	**0.400****	**0.077****	**0.195****	0.059**	0.049**
E*S*L	0.068**	0.017**	**0.086****	0.043**	0.049**	0.032**
E*S*T*L	0.011**	0.015**	0.008**	0.017**	0.004**	0.034**

* $p < 0.0167$; ** $p < 0.001$
T = test type; L = test length
E = estimate type; S = theta score level

Standard error. There is a significant main effect for test type for both IRT models. Under each model, the CAT (M_R = 0.285; M_{2pl} = 0.214) produces lower mean SEs than the CT (M_R = 0.379; M_{2pl} = 0.288). There is also a significant main effect for test length for both IRT models. Longer tests (M_R = 0.274; M_{2pl} = 0.215) produce lower SEs than shorter tests (M_R = 0.274; M_{2pl} = 0.215). For the 2PL model, there is a significant interaction between test type and test length; longer CATs produce lower mean SEs than shorter CATs or the CTs. A longer CT

did produce lower mean SEs than a shorter one. A main effect also exists for estimate type under both IRT models. EAP Normal estimate (M_R = 0.315; M_{2pl} = 0.239) result in lower SEs than EAP Uniform estimates (M_R = 0.349; M_{2pl} = 0.263). A significant interaction between test type and estimate type exists for both IRT models. Table 4.17 shows that CATs, in conjunction with EAP Normal, produce lower mean SEs for both IRT models. For the Rasch model, there is also a significant interaction between estimate type and test length (Table 4.18). Longer CATs using EAP Normal produce the lowest mean SEs. For the less informative prior (EAP Uniform), SE is more sensitive to the lack of information in the data and the prior has more of an influence, resulting in higher SEs.

Table 4. 17. *Mean SEs for the interaction effect of test type and estimate type for both IRT models, Simulation One, equally represented theta score levels.*

	Rasch Model		2PL Model	
	CAT	CT	CAT	CT
EAP Normal	0.275	0.354	0.211	0.267
EAP Uniform	0.295	0.404	0.217	0.309

Table 4. 18. *Mean SEs for the interaction between estimate type and test length for the Rasch model, Simulation One, equally represented theta score levels.*.

	15	30
EAP Normal	0.365	0.264
EAP Uniform	0.415	0.284

As expected, a significant main effect exists for theta score level. A U-shaped curve results when examining SE across all theta score levels. Mean SEs are higher at the extremes of the ability continuum and lower towards the middle (Figure 4.5). There is a significant interaction between theta score level and test type for both IRT models (Figure 4.6). CATs

produce lower mean SEs across all levels of the ability continuum. While the U-shaped pattern still exists, it is flatter for the CATs.

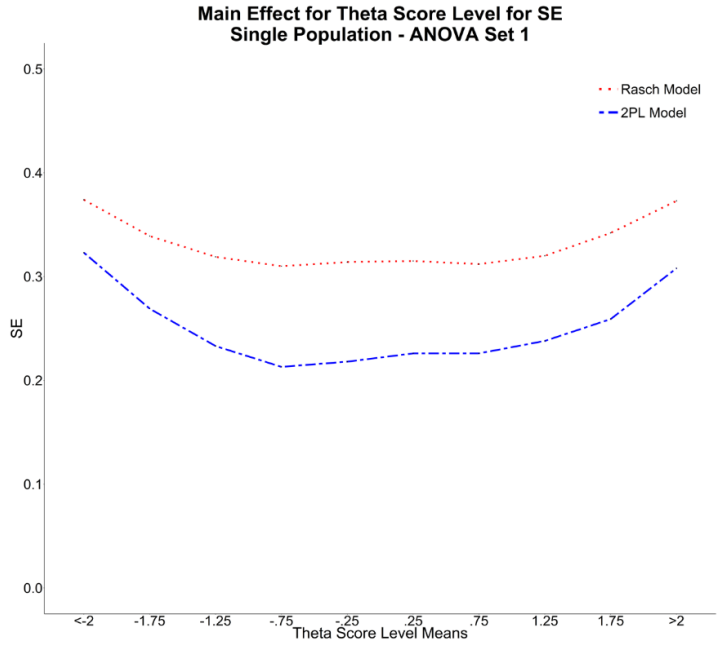

Figure 4. 5. Mean SEs for the main effect of theta score level for both IRT models for Simulation One, equally represented theta score levels.

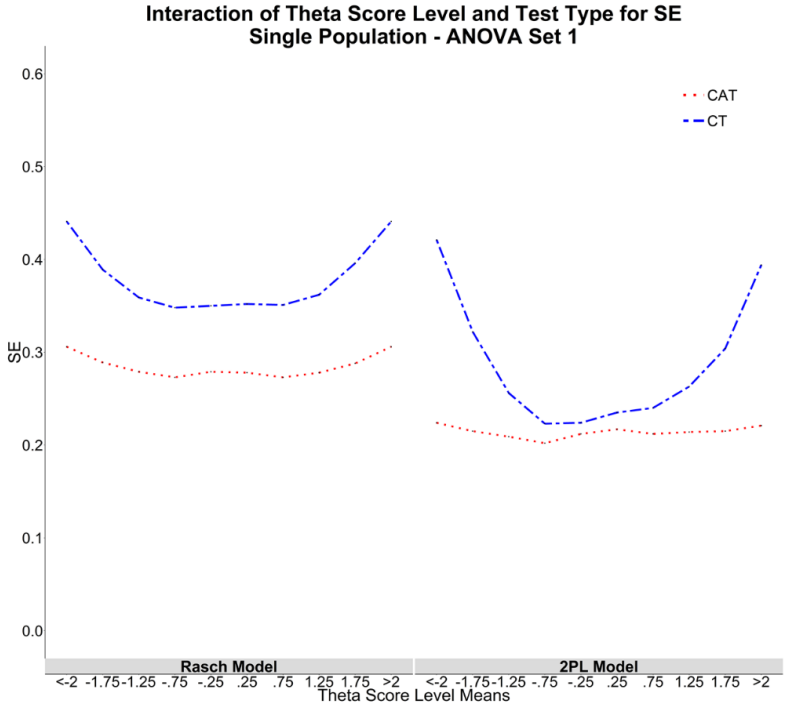

Figure 4. 6. Mean SEs for the interaction between theta score level and test type under both IRT models for Simulation One, equally represented theta score levels.

A two-way interaction between theta score level and estimate type is significant (Figure 4.7). EAP Normal estimates produce lower mean SEs across the ability continuum. However, the two estimate types are similar towards the middle of the ability spectrum. Lastly, a three-way interaction between theta score level, test type, and estimate type exists for both IRT models (Figure 4.8). When using a CAT, the differences between the two estimate types are negligible. Lower mean SEs still result for the CT when using EAP Normal.

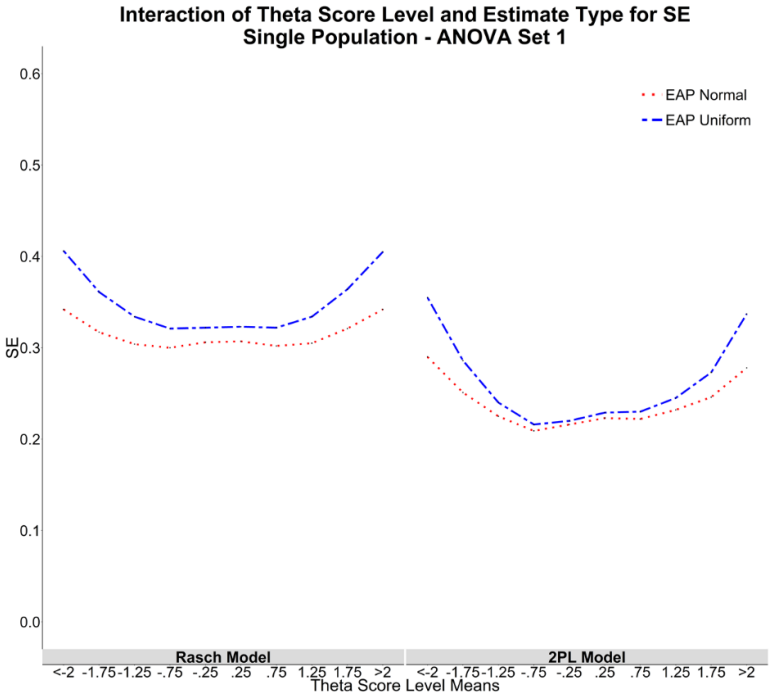

Figure 4. 7. Mean SEs for the interaction between theta score level and estimate type for under both IRT models for Simulation One, equally represented theta score levels.

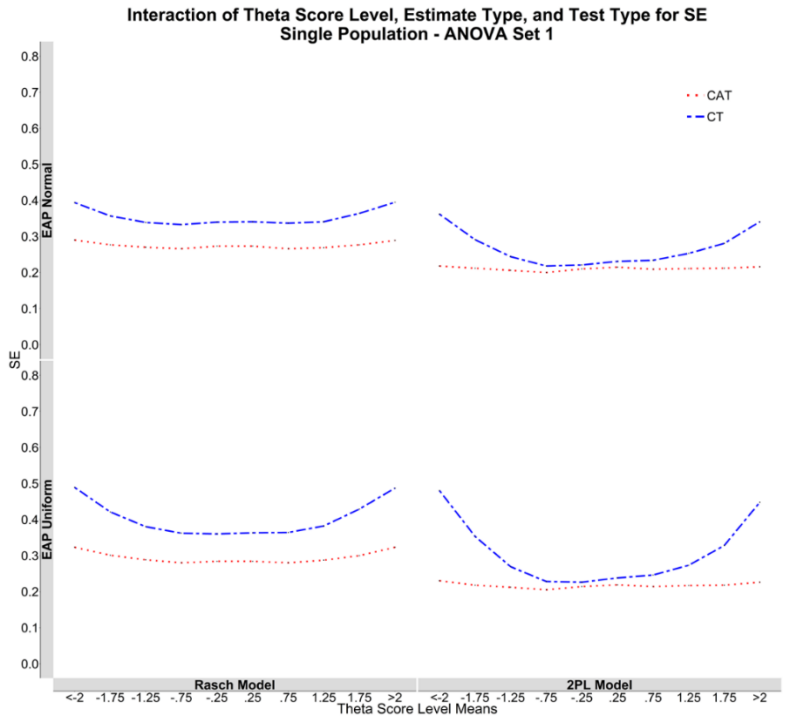

Figure 4. 8. Mean SEs for the interaction between theta score level, estimate type, and test type under both IRT models for Simulation One, equally represented theta score levels.

Bias. While no main effects exist for the between-subject factor, which is expected since bias usually cancels out across the whole ability spectrum, there is a significant interaction between test type and test length for the 2PL model. Table 4.19 shows that, while bias is similar for a CAT regardless of length, shorter CTs underestimate ability and longer CTs overestimate ability.

Table 4. 19. *Mean bias for the interaction between test type and test length for under the 2PL model for Simulation One, equally represented theta score levels.*

	CAT	CT
15	-0.001	-0.020
30	-0.002	0.017

For the 2PL model, there is a significant main effect for estimate type. EAP Normal (*M* = 0.001) produces less biased estimates than EAP Uniform (*M* = -0.004). However, the

differences are minimal. Also, for this IRT model, there is a significant interaction between estimate type and test type (Table 4.20). Again, differences are minimal for CAT. For the CT, the EAP Normal estimate results in more bias, but the differences are also small (0.009).

Table 4. 20. *Mean bias for the interaction between estimate type and test type under the 2PL model for Simulation One, equally represented theta score levels.*

	CAT	CT
EAP Normal	-0.002	0.003
EAP Uniform	-0.001	-0.006

Concerning theta score levels, there is a significant main effect and two significant interactions for both IRT models. Figure 4.9 shows the main effect for theta score level. A backward S-shaped pattern emerges. Estimates for lower ability simulees are often overestimated, represented by position bias. However, estimates for higher ability simulees are underestimated, represented by negative bias. This is as expected due to the influence of the prior; estimates are often pulled toward the prior's mean.

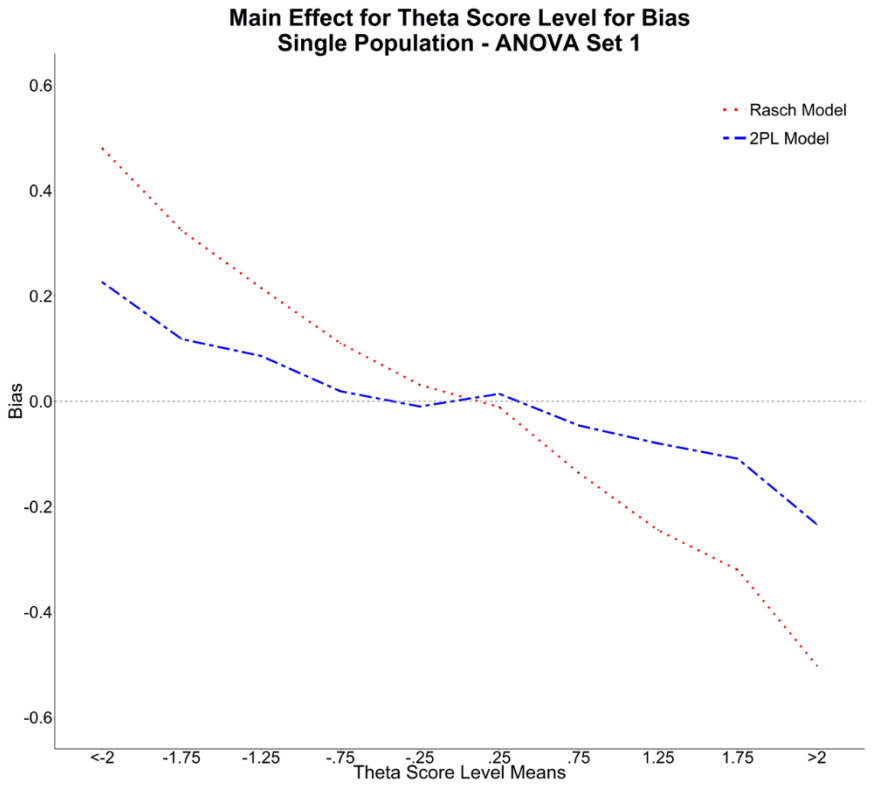

Figure 4. 9. Mean bias for the main effect of theta score level under both IRT models for Simulation One, equally represented theta score levels.

There is a significant interaction between theta score level and estimate type. As shown in Figure 4.10, EAP Uniform estimates generally flatten out the pattern observed for bias. Thus, the less informative prior has less of an effect and estimates are less influenced by the prior. There is also a significant three-way interaction between theta score level, estimate type, and test type (Figure 4.11). CATs (red) generally produce lower levels of bias under both IRT models, and EAP Uniform (bottom panel) produces less variability in bias across the entire ability continuum. Lastly, there is a significant interaction between theta score level, estimate type, and test length (not pictured). Longer tests produce less bias for EAP Normal estimates; bias levels are similar for EAP Uniform.

Figure 4. 10. Mean bias for the interaction between theta score level and estimate type under both IRT models for Simulation One, equally represented theta score levels.

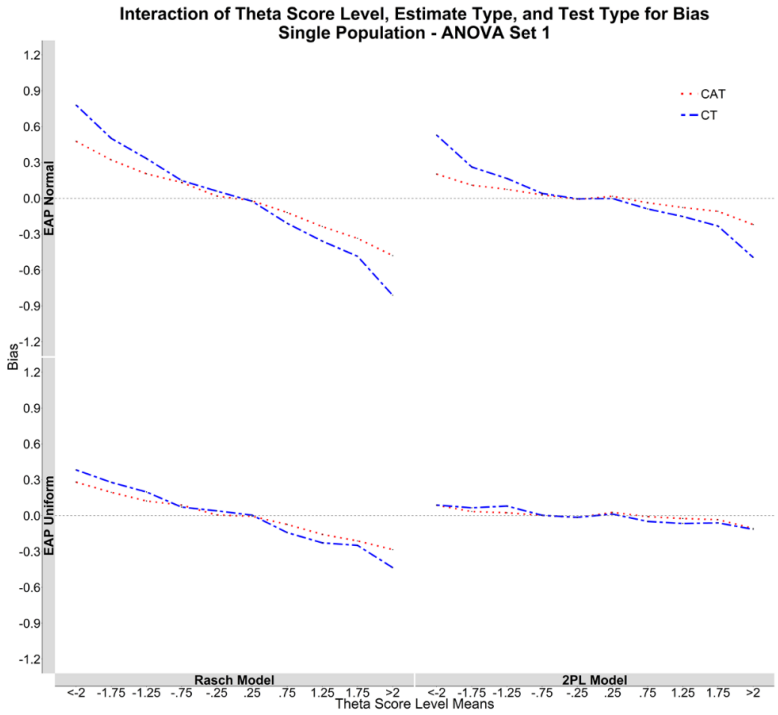

Figure 4. 11. Mean bias for the interaction between theta score level, estimate type, and test type under both IRT models for Simulation One, equally represented theta score levels.

RMSE. CATs (M_R = 0.450; M_{2pl} = 0.301) produce more accurate ability estimates than CTs (M_R = 0.543; M_{2pl} = 0.385) for both IRT models. This result was expected. Also, as expected, longer tests (M_R = 0.431; M_{2pl} = 0.296) produce more accurate estimates than shorter tests (M_R = 0.562; M_{2pl} = 0.390) for both IRT models. For the Rasch model, there is a significant main effect of estimate type. EAP Uniform (M = 0.485) produces more accurate ability estimates than EAP Normal (M = 0.508).

For the 2PL model, there are several significant interactions with estimate type. First, there is a significant interaction between estimate type and test length (Table 4.21). The two estimate types are similar for CAT, but EAP Normal estimates produce more accurate results for the CT. There is also a significant interaction between estimate type and test length (Table 4.22). The estimates are relatively similar for longer assessments and result in higher accuracy. Table 4.18 shows the mean RMSEs for the interaction between estimate type and test type for under the 2PL model. There is a significant interaction between estimate type, test type, and test length. As Figure 4.12 shows, estimate accuracy is highest when a long CAT is utilized; little difference exists between estimate type.

Table 4. 21. *Mean RMSE for the interaction between estimate type and test length for under the 2PL model for Simulation One, equally represented theta score levels.*

	15	30
EAP Normal	0.383	0.298
EAP Uniform	0.398	0.293

Table 4. 22. *Mean RMSE for the interaction between estimate type and test type for mean RMSE under the 2PL model for Simulation One, equally represented theta score levels.*

	CAT	CT
EAP Normal	0.302	0.379
EAP Uniform	0.300	0.391

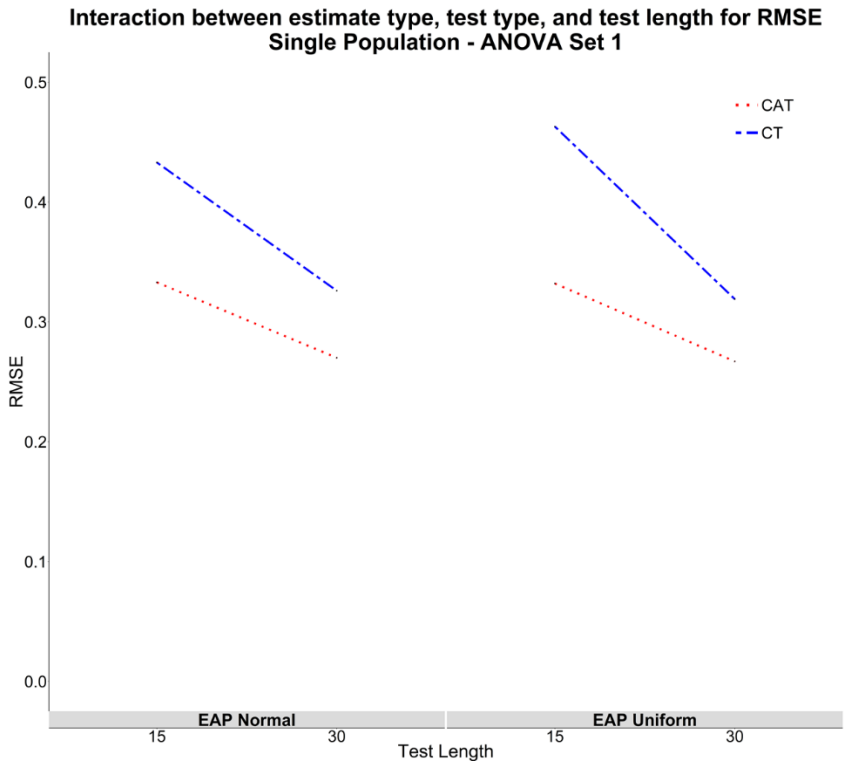

Figure 4. 12. Mean RMSE for the interaction between estimate type, test type, and test length for the 2PL model for Simulation One, equally represented theta score levels.

There is a significant main effect of theta score level for both IRT models (Figure 4.13). Just like with SE, a U-shaped pattern emerges. Simulees with true abilities near the middle of the ability continuum have more accurate estimates than those towards the extremes. A significant interaction between theta score level and test type occurs for both IRT models (Figure 4.14). CATs generally produce more accurate estimate across all theta levels than CTs, but accuracy differences are small towards the middle of the continuum. Lastly, there is a significant interaction between theta score level and estimate type for both IRT models (Figure 4.15). EAP Uniform estimates produce more accurate estimates at the high and low extremes of the theta continuum. Towards the middle, EAP Normal produces more accurate estimates but differences are minute.

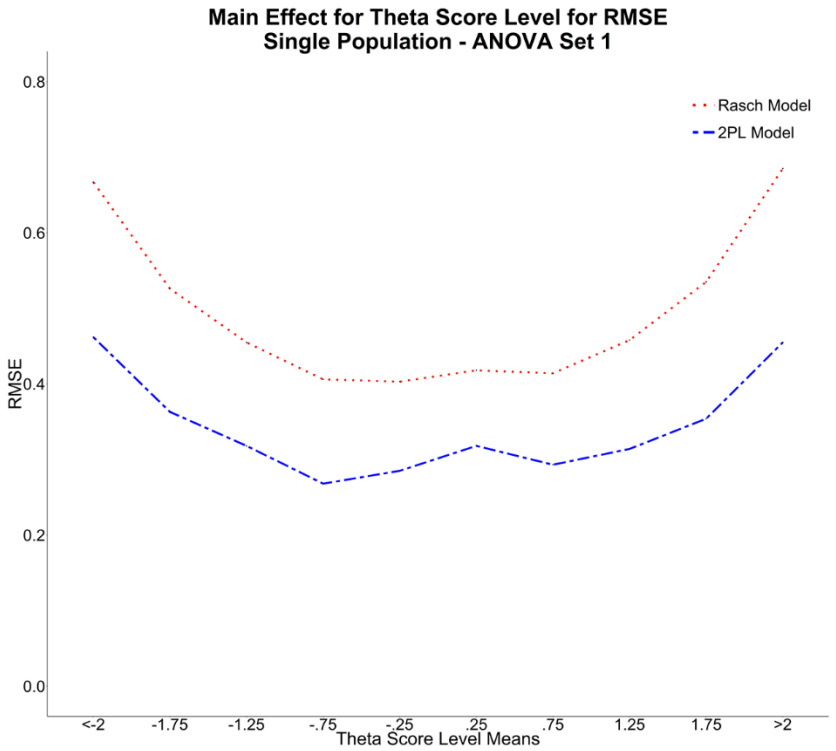

Figure 4. 13. Mean RMSE values for the main effect of theta score level under both IRT models for Simulation One, equally represented theta score levels.

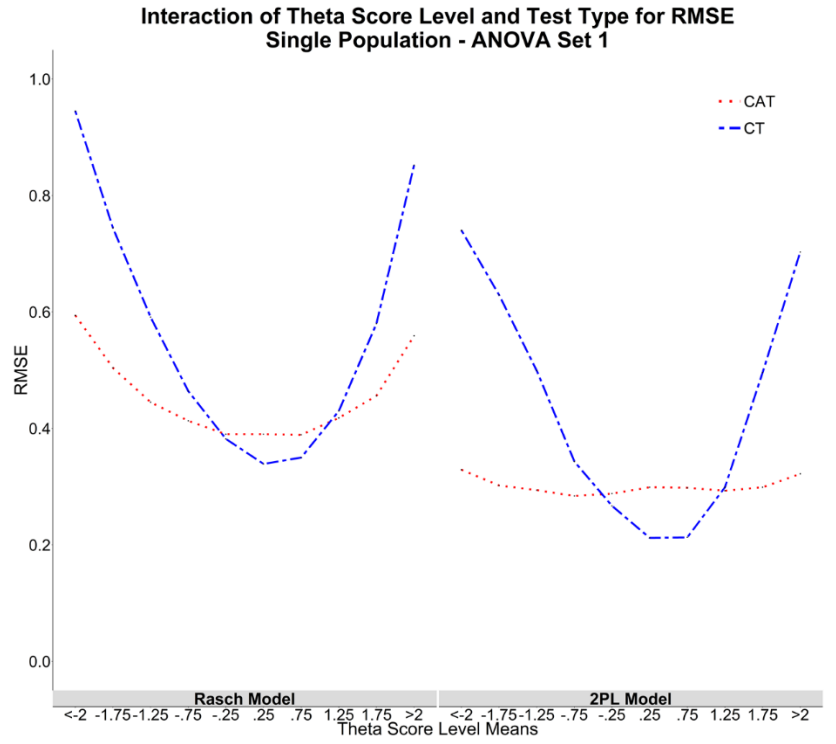

Figure 4. 14. Mean RMSEs for the interaction between theta score level and test type for mean RMSE under both IRT models for Simulation One, equally represented theta score levels.

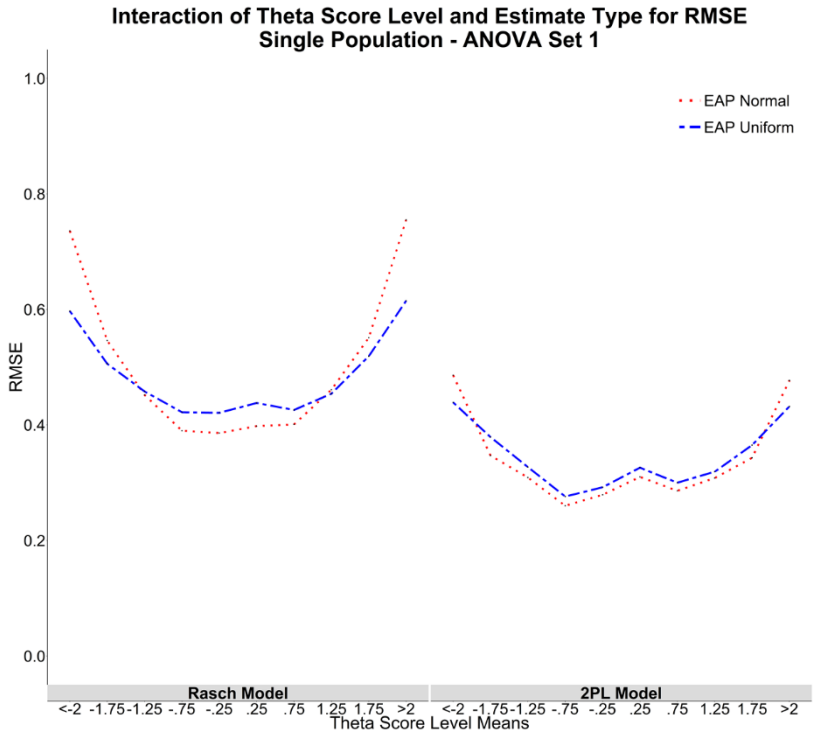

Figure 4. 15. Mean RMSEs for the interaction between theta score level and estimate type for mean RMSE under both IRT models for Simulation One, equally represented theta score levels.

ANOVA Set 2. The second set of ANOVAs presented deal with the peaked distribution for the Single Population. In other words, this sample set has a distribution that matches that seen in the population. While the number of factors involved in this analysis is small, it allows for a look at the how the estimate type, test type, and test length functions in a peaked population where the test prior mean matches the ability mean. Table 4.23 presents the η_w^2 for all effects and interactions for each dependent variable, separated by IRT model. There are very few meaningful results; most relate to standard error when using the Rasch model. This is because the item discrimination value in the Rasch model is 1, meaning all items are equally discriminating. However, for the 2PL model, items vary in discriminatory power. It is pertinent to note that only one effect concerning bias is significant, but does not meet the effect size cutoff.

Table 4. 23. *Within family effect sizes from ANOVA Set 2 conducted on standard error (SE), bias, root mean square error (RMSE) for each IRT model using the Single Population.*

Effect	SE		Bias		RMSE	
	Rasch	**2PL**	**Rasch**	**2PL**	**Rasch**	**2PL**
T	**0.214****	**0.226****	0.000	0.000	0.061**	**0.096****
L	**0.231****	0.024**	0.000	0.002	**0.136****	**0.110****
T*L	0.015**	0.010**	0.000	0.003**	0.001	0.016**
E	**0.546****	**0.300****	0.000	0.000	0.055**	0.002*
E*T	**0.097****	**0.158****	0.000	0.000	0.006**	0.003**
E*L	**0.105****	0.017**	0.000	0.000	0.002*	0.006**
E*T*L	0.007**	0.007**	0.000	0.000	0.005**	0.005**

* $p < 0.0167$; ** $p < 0.001$
T = test type; L = test length
E = estimate type

Standard error. For both IRT models, the CAT ($M_R = 0.285$; $M_{2pl} = 0.214$) produces lower mean SEs than the CT ($M_R = 0.379$; $M_{2pl} = 0.289$). For the Rasch model, there is a significant main effect for test length; longer tests ($M = 0.274$) produce lower SEs than shorter tests ($M = 0.390$). For both IRT models, there is a significant main effect for estimate type. EAP Normal ($M_R = 0.315$; $M_{2pl} = 0.240$) results in lower mean SEs than EAP Uniform ($M_R = 0.349$; $M_{2pl} = 0.264$). The interaction between estimate type and test type (Table 4.24) is significant for both IRT models and shows that CATs using EAP Normal produces the lowest SEs. A CT utilizing EAP Uniform produces the highest SEs. For the Rasch model, the interaction between estimate type and test length (Table 4.25) shows that longer assessment using EAP Normal estimate produce lower SEs.

Table 4. 24. *Mean SEs for the interaction between estimate type and test type for under both IRT models for Simulation One, peaked distributions.*

	Rasch Model		2PL Model	
	CAT	CT	CAT	CT
EAP Normal	0.275	0.354	0.211	0.268
EAP Uniform	0.295	0.403	0.218	0.309

Table 4. 25. *Mean SEs for the interaction between estimate type and test length for under the Rasch model for Simulation One, peaked distributions.*

	15	30
EAP Normal	0.365	0.265
EAP Uniform	0.414	0.284

RMSE. Few effects for RMSE are meaningful; only two meet the cutoff values. First, for the 2PL model, there is a significant main effect for test type. CATs (M = 0.295) produce more accurate estimates than CTs (M = 0.383). This result is expected. While this main effect did not occur for the Rasch model, the effect size was relatively high (M = 0.061). For models, there is a significant main effect for test length. As expected, longer tests (M_R = 0.424; M_{2pl} = 0.292) result in more accurate estimates than shorter tests (M_R = 0.561; M_{2pl} = 0.386).

Simulation Two

This section utilizes the Subgroup Population data, in which a majority group (A) and a minority group (B) both exist. The minority group is further subdivided into two subgroups, where one group has a lower ability (B_L) than the other (B_H). B_H is identical to A. However, B_H composes only 20% of the minority group (i.e., Group B) and 6% of the total sample (Groups A and B). However, to ensure that this small group size does not influence results, groups were constrained to be equal sizes for Sample Set 1. For Sample Set 3, however, sample sizes were smaller since the population distribution was approximated.

ANOVA Set 1. The first set of ANOVAs to be presented are those that utilized the uniform sample, where all theta score levels are equally represented. This approach is taken because when group membership is examined in conjunction with theta score level, the influence of test prior is unbalanced. First, information pertaining to the CTs is given, followed by CAT.

It was inappropriate to analyze both test types together, since the CT had four final estimate priors, while the CATs had three test priors.

CT. The CTs did not involve priors during test administration. Prior information is only influential during final ability estimation after the test has been administered. Thus, the utilized priors are referred to as final estimate priors. Table 4.26 presents the effect sizes for this portion of the study. The four test priors for the CTs are: Population Composite Prior, Group Composite Prior, Group Specific Prior, and Less Informative Prior.

Table 4. 26. *Effect sizes from CT ANOVA Set 1 conducted on standard error (SE), bias, root mean square error (RMSE) for each IRT model using the Subgroup Population.*

Effect	SE		Bias		RMSE	
	Rasch	**2PL**	**Rasch**	**2PL**	**Rasch**	**2PL**
Results on Final Estimate Priors						
P	**0.203****	**0.422****	**0.230****	**0.455****	**0.140****	0.067**
P*S	**0.108****	**0.104****	0.043**	**0.226****	0.035**	**0.076****
P*S*L	0.015**	0.005**	0.005**	0.013**	0.003**	0.013**
P*L	0.039**	0.028**	0.057**	**0.074****	0.001	0.015**
Results Dealing with All Other Factors						
L	**0.750****	**0.524****	0.006*	0.032**	**0.617****	**0.707****
S	**0.771****	**0.824****	**0.901****	**0.658****	**0.814****	**0.766****
S*L	0.028**	0.010**	0.005**	0.010**	0.010**	0.006**

* $p < 0.0167$; ** $p < 0.001$
P = final estimate prior; L = test length
S = theta score level

Pertinent final estimate prior results. First presented are the results that relate to the final estimate priors. Any main effects and interaction that are significant and meet the effect size criterion are presented.

Standard error. There is a significant main effect of final estimate prior on mean SE. Table 4.27 presents the mean SE for each prior. As can be seen, when a Less Informative Prior is used to estimate ability, there is more error obtained within the estimate. There is a significant

interaction between final estimate prior and theta score level (Figure 4.16). When there is a Less Informative Prior placed on the estimate (EAP Uniform for the CT), the mean SE gets quite high, especially at the ends of the theta continuum. This is because the mean of the test is far from these theta levels and less information is obtained; therefore, estimation suffers.

Table 4. 27. *Mean SEs for the main effect of final estimate prior under both IRT models for Simulation Two, CT, equally represented theta score levels*.

	Rasch Model	2PL Model
Population Composite Prior	0.335	0.279
Group Composite Prior	0.338	0.284
Group Specific Prior	0.337	0.283
Less Informative Prior	0.401	0.384

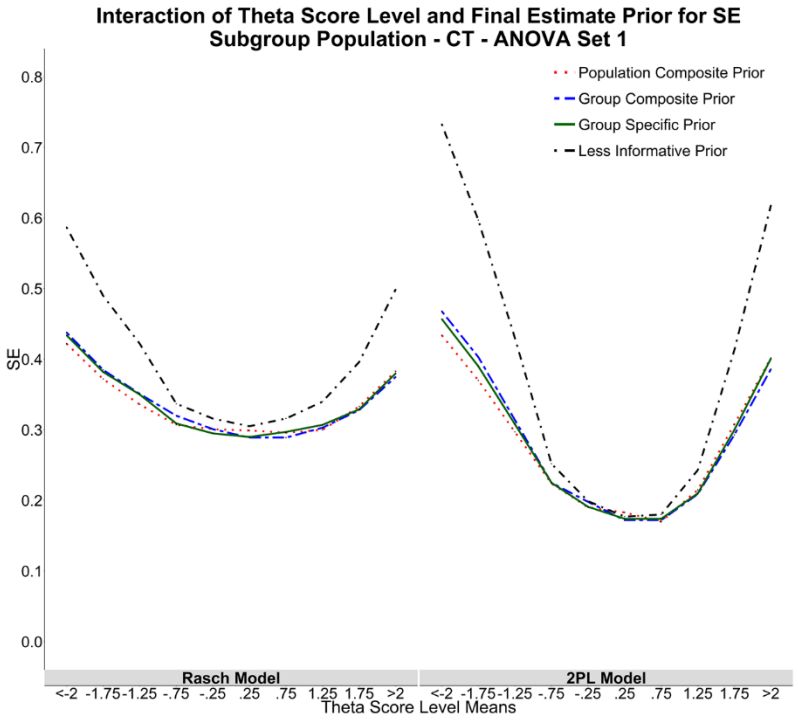

Figure 4. 16. Mean SEs for the interaction between theta score level and final estimate prior under both IRT models for Simulation Two, CT, equally represented theta score levels.

Bias. There is a significant main effect of final estimate prior when examining bias as well. Table 4.28 presents the mean bias, and absolute bias, for the four final estimate priors

under each IRT model. While SE may be high for the Less Informative Prior condition, it has the lowest bias out of the four test priors.

Table 4. 28. *Mean bias for the main effect of final estimate prior for mean bias, with absolute bias, for both IRT models for Simulation Two, CT, equally represented theta score levels.*

	Rasch Model		2PL Model	
	Bias	Absolute Bias	Bias	Absolute Bias
Population Composite Prior	0.112	0.112	0.075	0.075
Group Composite Prior	0.071	0.071	0.039	0.039
Group Specific Prior	0.095	0.095	0.057	0.057
Less Informative Prior	0.051	0.051	-0.022	0.022

There is a significant interaction between theta score level and final estimate prior for bias for the 2PL model. The three informative priors are similar in terms of mean bias (Figure 4.17), with more bias at the extremes of the ability continuum. However, although using a less informative prior (i.e., Less Informative Prior, black line) results in higher SEs, these estimates have lower levels of bias across the entire ability spectrum.

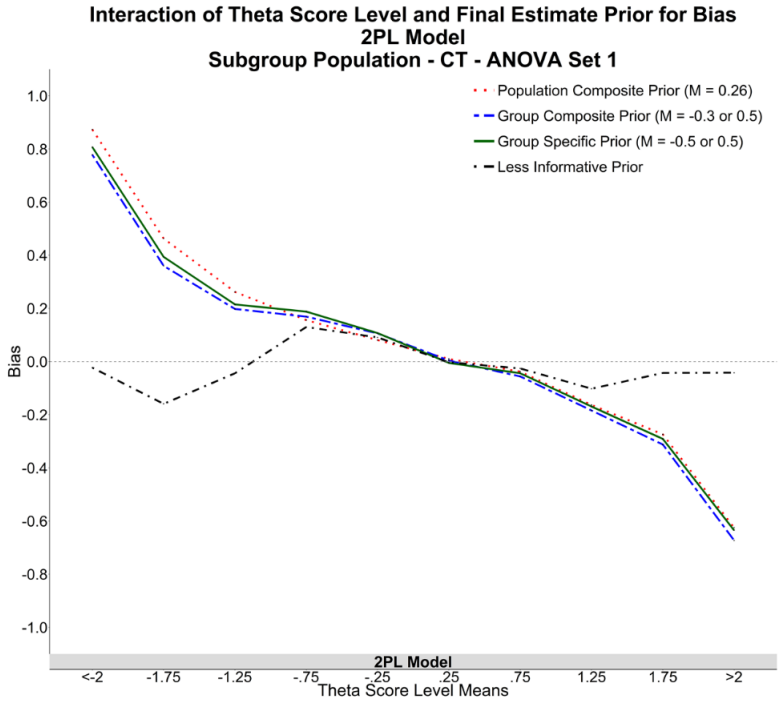

Figure 4. 17. Mean bias for the interaction between theta score level and final estimate prior under both IRT models for Simulation Two, CT, equally represented theta score levels.

There is a significant interaction between final estimate prior and test length, presented in Table 4.29, for the 2PL model. Longer tests help when using certain priors (i.e., Population Composite Prior, Less Informative Prior) and decrease bias, but with other test priors (i.e., Group Composite Prior, Group Specific Prior), bias increases.

Table 4. 29. *Mean bias for the interaction between final estimate prior and test length under the 2PL model for Simulation Two, CT, equally represented theta score levels.*

	15	30
Population Composite Prior	0.085	0.064
Group Composite Prior	0.027	0.051
Group Specific Prior	0.051	0.064
Less Informative Prior	-0.053	0.009

RMSE. For mean RMSE, there is a significant main effect of final estimate prior for the Rasch model (Table 4.30). Using an EAP estimate with a less informative prior (i.e., Less Informative Prior) leads to more accurate ability estimates. For the 2PL model, the effect size was not reached but was close. There is also an interaction between final estimate prior and theta score level (Figure 4.18) for the 2PL model. The final estimate priors have similar levels of accuracy towards the middle of the ability continuum. However, using a less informative prior led to more accurate estimates at the extremes.

Table 4. 30. *Mean RMSE for the main effect of final estimate prior for the Rasch model for Simulation Two, CT, equally represented theta score levels.*

	Rasch Model
Population Composite Prior	0.599
Group Composite Prior	0.590
Group Specific Prior	0.593
Less Informative Prior	0.540

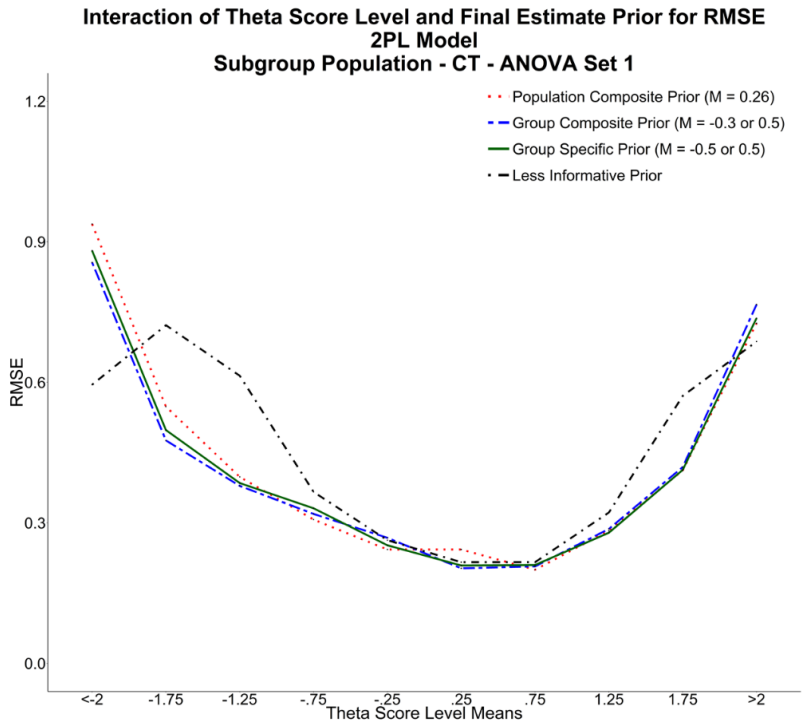

Figure 4. 18. Mean bias for the interaction between test prior and theta score level for the 2PL model for Simulation Two, CT, equally represented theta score levels.

Other meaningful results. All other significant results meeting the effect size criterion are presented.

Standard error. A significant main effect for test length exists for SE. Longer tests (M_R = 0.299; M_{2pl} = 0.258) produces lower levels of standard error than shorter tests (M_R = 0.406; M_{2pl} = 0.357). There is a significant main effect for theta score level (Figure 4.19). The resulting pattern is U-shaped; SE means are higher for the extreme ability levels than for the middle ones. This is expected when ignoring other influences, as seen in Simulation One.

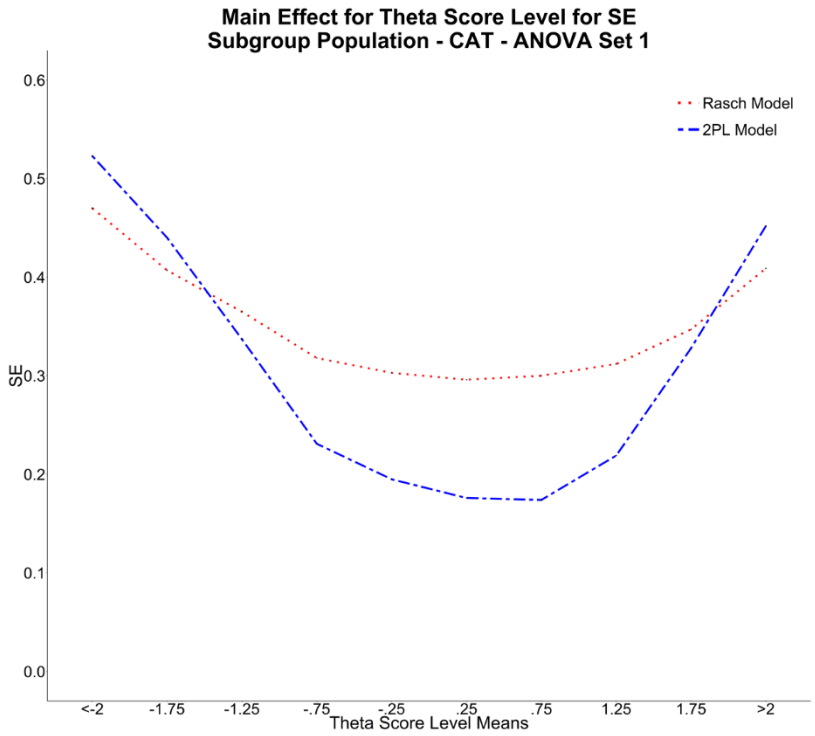

Figure 4. 19. Mean SE for the main effect of theta score level for both IRT models in Simulation Two, CT, equally represented theta score levels.

Bias. There is a significant main effect of theta score level (Figure 4.20) for bias. As can been seen, a S-shaped curve results, where lower ability levels are overestimated (i.e., positive bias) and higher ability levels are underestimated (i.e., negative bias). When test length is added in for an interaction, longer tests generally result in less bias.

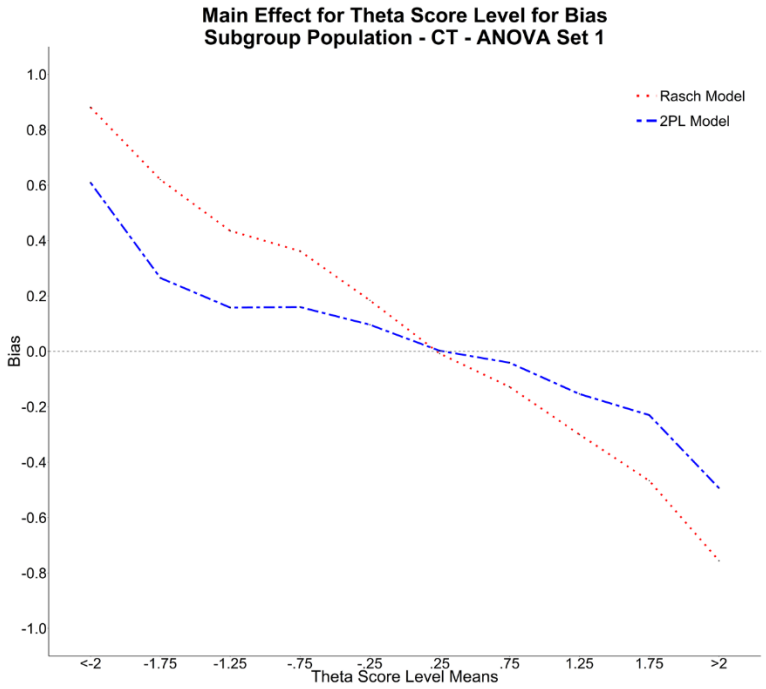

Figure 4. 20. Mean bias for the main effect of theta score level for both IRT models for Simulation Two, CT, equally represented theta score levels.

RMSE. For mean RMSE, there are two significant main effects – test length and theta score level. Longer tests (M_R = 0.531; M_{2pl} = 0.380) results in more accurate estimates than shorter tests (M_R = 0.630; M_{2pl} = 0.483). Figure 4.21 shows the main effect for theta score level. As expected, there is a U-shaped pattern to RMSE for both IRT models. For simulees at the extremes of the latent continuum, ability estimates are less accurate than for those near the middle of the continuum.

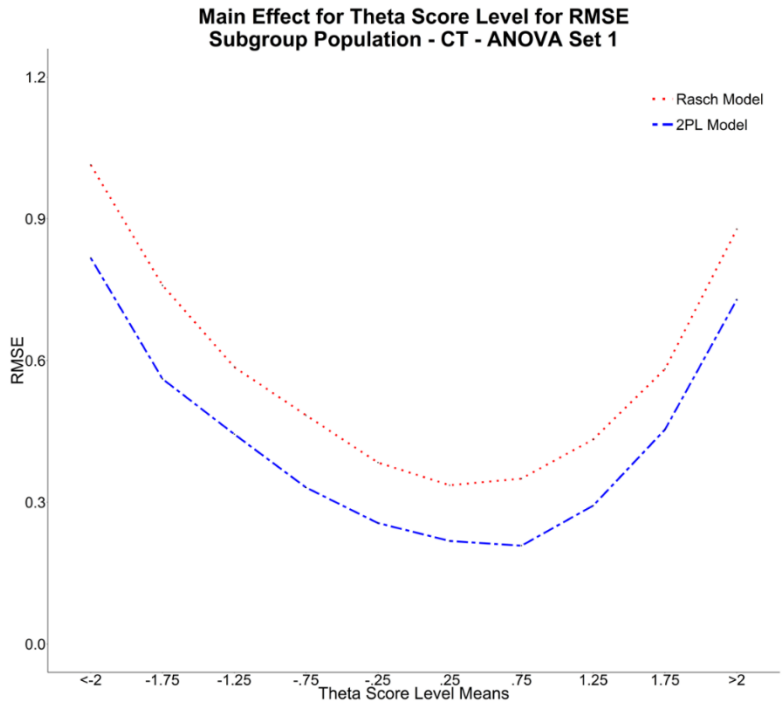

Figure 4. 21. Mean RMSE for the main effect of theta score level under both IRT models for Simulation Two, CT, equally represented theta score levels.

CAT. For the CATs, test priors are used throughout test administration and the prior used in EAP Normal estimation. Table 4.31 presents the η_w^2 for all effects and interactions for each dependent variable, separated by IRT model. Effects involving the test priors will be presented first, followed by all other meaningful effects.

Table 4. 31. *Within family effect sizes from CAT ANOVA Set 1 conducted on standard error (SE), bias, root mean square error (RMSE) for each IRT model using the Subgroup Population.*

Effect	SE		Bias		RMSE	
	Rasch	**2PL**	**Rasch**	**2PL**	**Rasch**	**2PL**
Results on Test Priors						
P	0.000	0.000	**0.205****	0.031	0.001	0.001
P*S	0.018**	0.016**	0.001*	0.002	0.011**	0.010**
P*S*L	0.003**	0.001	0.000*	0.003	0.002	0.002
P*E*S	0.014**	0.000**	0.000**	0.000	0.020**	0.008**
P*E*S*L	0.007**	0.001	0.000	0.000	0.004**	0.001
P*L	0.000	0.000	0.007	0.001**	0.000	0.001
P*E	0.000*	0.000**	**0.341****	**0.364****	0.007**	0.006
P*L*E	0.000	0.000**	0.040**	0.012**	0.004*	0.000
Results Dealing with All Other Factors						
E	**0.759****	**0.909****	**0.430****	**0.376****	**0.767****	**0.071****
L	**0.998****	**0.952****	0.005	0.012*	**0.837****	**0.611****
E*L	**0.239****	**0.084****	0.062**	0.017**	0.040**	0.035**
S	**0.638****	**0.195****	**0.753****	**0.397****	**0.344****	**0.034****
S*L	**0.116****	0.023**	0.041**	0.004**	0.040**	0.004*
E*S	**0.608****	**0.545****	**0.875****	**0.944****	**0.649****	**0.395****
E*S*L	**0.239****	0.061**	0.101**	0.035**	0.082**	0.004**

* $p < 0.0167$; ** $p < 0.001$
L = test length; P = test prior
E = estimate type; S = theta score level

Pertinent test prior results. First, the results relating to the test prior are presented, since these are the primary effects of interest. When examining theta score levels in relation to standard error, there is no effect of test prior in comparison to the means of the theta score levels.

Bias. There is a significant main effect for test prior when examining bias for the Rasch model. The Population Composite Prior (M = 0.046) produces the highest level of bias, followed by the Group Specific Prior (M = 0.026). The Group Composite Prior (M = -0.003) produces the lowest amount of bias. For both IRT models, there is a significant interaction between test prior and estimate type. The test prior is used during the CAT for both ability initialization and for item selection. For estimate type, EAP Normal utilizes the test prior used during the test while

EAP Uniform has a less informative prior. Although the test prior is used during test administration, less information is being used from the prior to obtain final estimates of ability. Table 4.32 shows the mean bias for the interaction of these two factors. The Group Composite Prior tends to underestimate ability, whereas the other two priors tend to overestimate ability. The Population Composite Prior produces the most bias. However, bias decreases when the EAP Uniform estimate is used.

Table 4. 32. *Means for bias when examining the interaction between test prior and final estimate type for both IRT models for Simulation Two, CAT, equally represented theta score levels.*

	Rasch Model		2PL Model	
	EAP Normal	EAP Uniform	EAP Normal	EAP Uniform
Population Composite Prior	0.056	0.035	0.014	0.003
Group Composite Prior	-0.005	-0.002	-0.006	-0.004
Group Specific Prior	0.033	0.018	0.008	0.001

Other meaningful results. The remaining results are presented next. Most of these results were as expected based off the results from the base simulation, Simulation One.

Standard error. There is a significant main effect of estimate type for both IRT models for SE. EAP Normal (M_R = 0.275; M_{2pl} = 0.211) produces lower mean SEs than EAP Uniform (M_R = 0.296; M_{2pl} = 0.218). A significant main effect for test length also occurs for both IRT models. Longer tests (M_R = 0.234; M_{2pl} = 0.195) produces lower mean SEs than shorter tests (M_R = 0.337; M_{2pl} = 0.234). There is also a significant interaction between estimate type and test length (Table 4.33). Longer assessments utilizing EAP Normal estimation produce lower mean SEs.

Table 4. 33. *Mean SEs for the interaction between estimate type and test length for both IRT models for Simulation Two, CAT, equally represented theta score levels.*

	Rasch Model		2PL Model	
	15	30	15	30
EAP Normal	0.322	0.229	0.230	0.193
EAP Uniform	0.353	0.238	0.238	0.197

A significant main effect, as expected, exists for theta score level for SE. As Figure 4.22 shows, mean SE is higher at the extremes of the ability continuum. However, the differences are minimal. There is a significant two-way interaction for theta score level and test length for the Rasch model, in which longer tests reduces SE, as well as a significant two-way interaction for theta score level and estimate type for both IRT models, which is presented in Figure 4.23. EAP Uniform estimates produce higher mean SEs at all levels of the ability continuum, and there is an upward turn at the extremes for both final estimate type. There is a significant three-way interaction for the Rasch model including test length (not pictured). A longer test produced lower levels of mean SE.

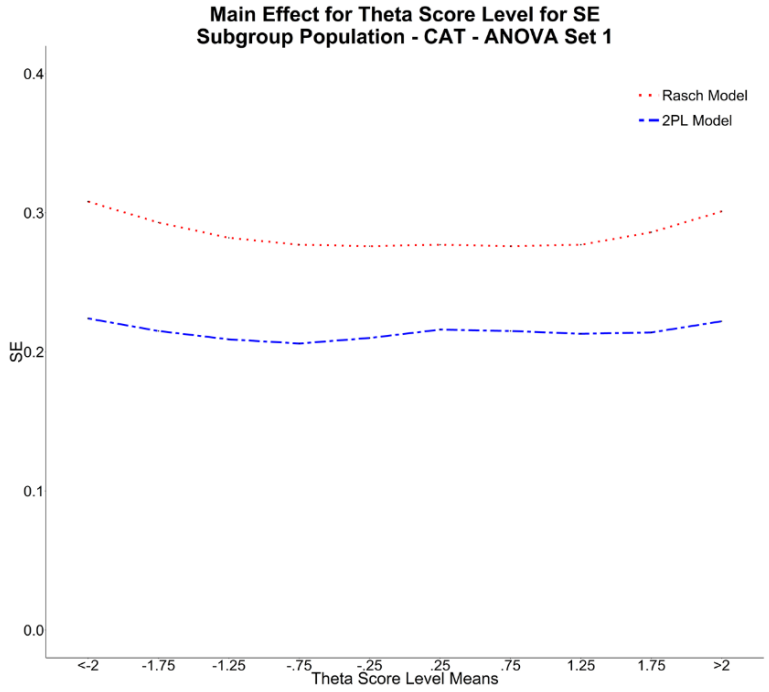

Figure 4. 22. Mean SE for the main effect for theta score level under both IRT models for Simulation Two, CAT, equally represented theta score level.

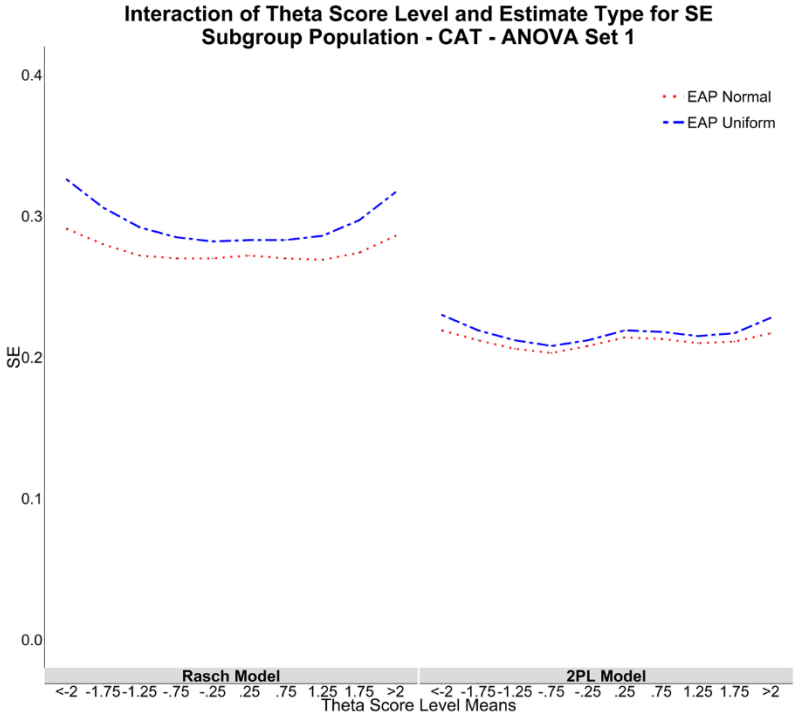

Figure 4. 23. Mean SEs for the interaction between theta score level and estimate type under both IRT models for Simulation Two, CAT, equally represented theta score levels.

Bias. For bias, there is a significant main effect for estimate type for both IRT models. EAP Uniform (M_R = 0.017; M_{2pl} = 0.000) produces lower levels of mean bias than EAP Normal (M_R = 0.028; M_{2pl} = 0.005). There is also a significant main effect for theta score level. As previously seen, a backwards S-shape pattern occurs (Figure 4.24). The lower ability levels are generally overestimated, where the higher ability levels are underestimated. A two-way interaction between theta score level and estimate type is significant (Figure 4.25). Using EAP Uniform (blue) for final ability estimation produces lower levels of bias across the ability spectrum than using EAP Normal (red).

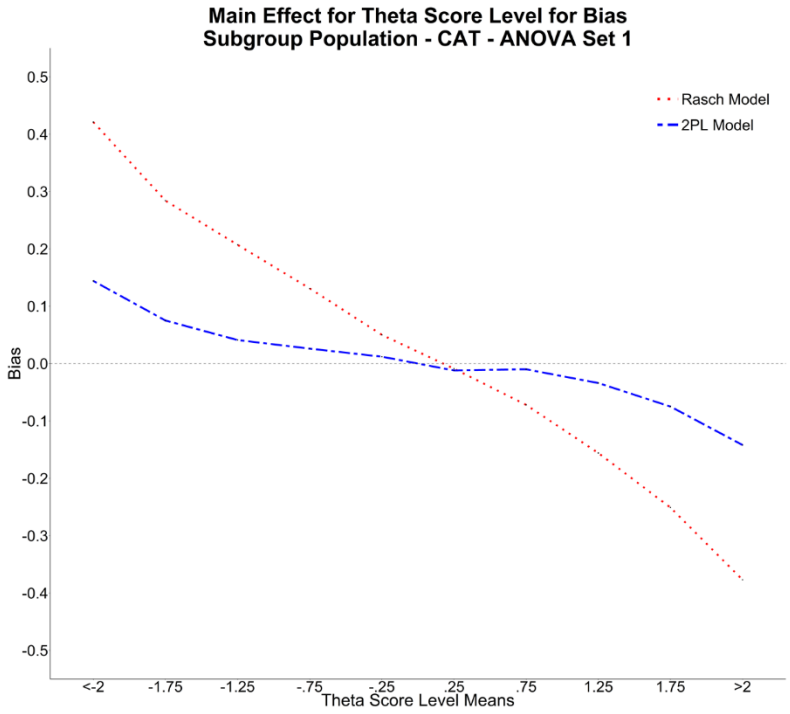

Figure 4. 24. Mean bias for the main effect of theta score level under both IRT models for Simulation Two, CAT, equally represented theta score levels.

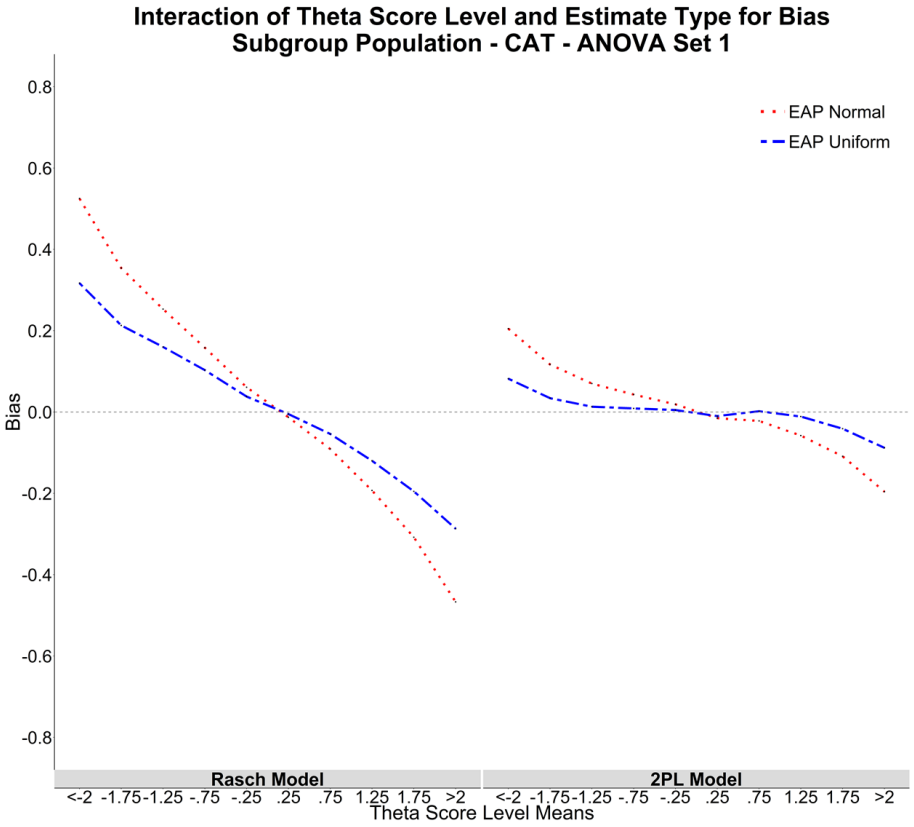

Figure 4. 25. Mean bias for the two-way interaction (theta score level x estimate type) under both IRT models for Simulation Two, CAT, equally represented theta score levels.

RMSE. There is a significant main effect for estimate type for mean RMSE for both IRT models. EAP Uniform (M_R = 0.445; M_{2pl} = 0.300) produces slightly more accurate estimates than EAP Normal (M_R = 0.465; M_{2pl} = 0.302). There is also a significant main effect of test length for both IRT models. Longer assessments (M_R = 0381; M_{2pl} = 0.272) produce more accurate ability estimates than shorter assessments (M_R = 0.529; M_{2pl} = 0.330).

Also, for both IRT models, there is a significant main effect of theta score level (Figure 4.26). Estimates are more accurate towards the middle of the ability continuum. A significant two-way interaction (theta score level by estimate type) for both IRT models shows that EAP Uniform estimates are more accurate towards the extremes of the latent trait continuum. This is because the less informative estimate type and its prior is more sensitive to a lack of information

at these ability levels and, thus, has more of an influence over the ability estimates. However, EAP Normal is more accurate towards the middle, but the two estimate types are close (Figure 4.27).

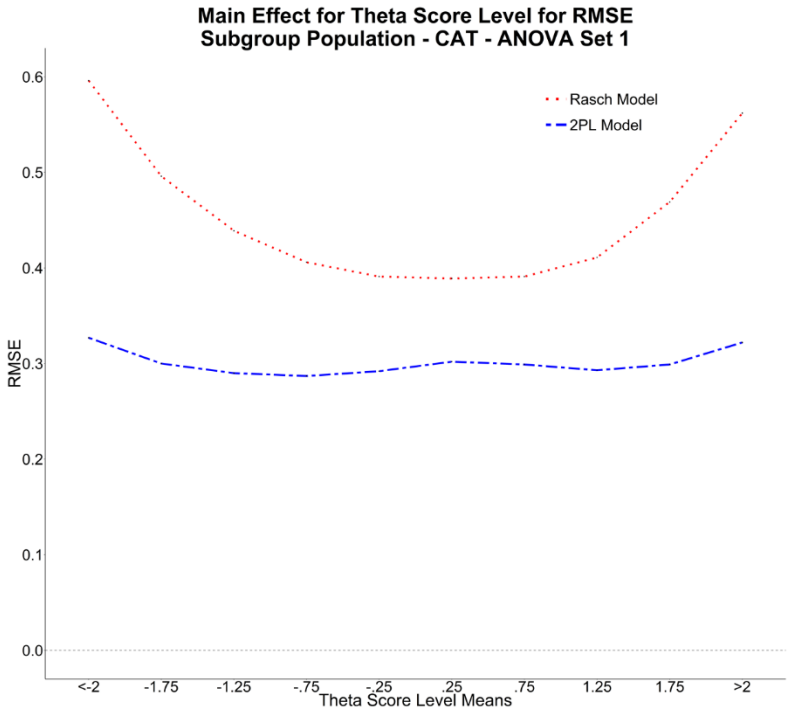

Figure 4. 26. Mean RMSE for the main effect of theta score level under both IRT models for Simulation Two, CAT, equally represented theta score levels.

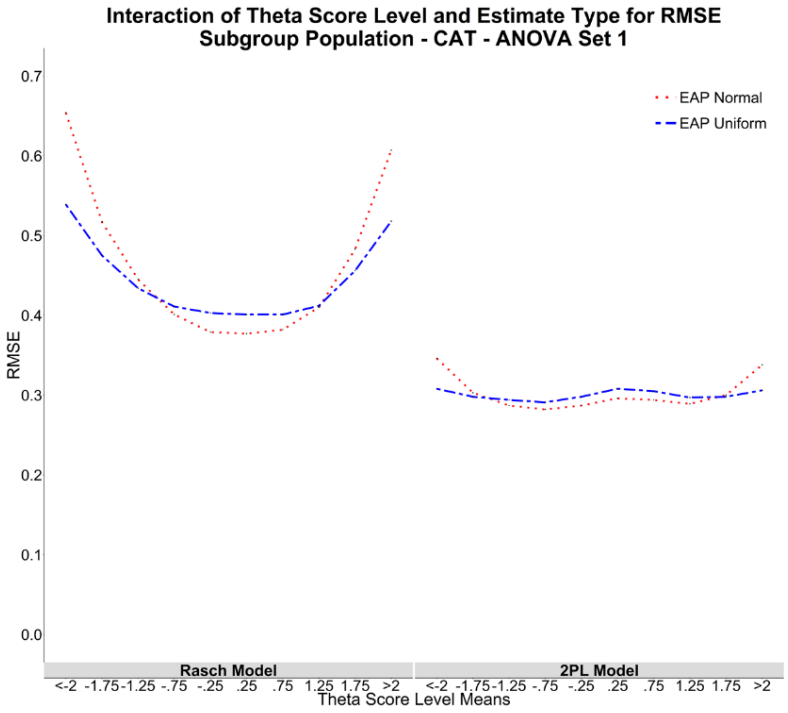

Figure 4. 27. Mean bias for the interaction of theta score level and estimate type under both IRT models for Simulation Two, CAT, equally represented theta score levels.

ANOVA Set 2. The second set of ANOVAs to be presented are those that utilized the peaked sample, where the subgroup distributions approximated the means observed in the population. This allowed the interaction of test prior and group membership to be observed without any erroneous influence with other factors, such as theta score level. As with previous section, results are first presented for CTs then CATs.

CT. While ANOVA Set 1 examined equal theta score levels, ANOVA Set 2 examined the groups with distributions approximating the population. Table 4.34 presents the effect size for all effects and interactions. Again, priors influence only final ability estimation, and are called final estimate priors.

Table 4. 34. *Within-family effect sizes from CT ANOVA Set 2 conducted on standard error (SE), bias, root mean square error (RMSE) for each IRT model using the Subgroup Population.*

Effect	SE		Bias		RMSE	
	Rasch	**2PL**	**Rasch**	**2PL**	**Rasch**	**2PL**
Results on Final Estimate Priors						
P	**0.154****	**0.446****	**0.204****	**0.263****	**0.071****	**0.217****
P*G	**0.130****	**0.196****	0.040**	**0.305****	0.069**	**0.123****
P*G*L	0.033**	0.015**	0.006**	0.023**	0.011**	0.011**
P*L	0.000**	0.000**	0.042**	0.079**	0.009**	0.029**
Results Dealing with All Other Factors						
L	**0.808****	**0.522****	**0.071****	**0.111****	**0.780****	**0.659****
G	**0.687****	**0.670****	**0.920****	**0.583****	**0.506****	**0.460****
G*L	0.040**	0.011**	0.000	0.000	0.003*	0.002*

* $p < 0.0167$; ** $p < 0.001$
P = final estimate prior; L = test length
G = group membership

Pertinent test prior results. First, any results relating to the final estimate priors are presented.

Standard error. A significant main effect exists for final estimate prior when examining mean SE (Table 4.35). Mean SEs are similar for all final estimate priors except for the Less Informative Prior condition. These SEs are higher due to the difficulty of estimating abilities at the extremes. There is a significant interaction between final estimation prior and group membership for both models (Figure 4.28). The Less Informative Prior condition has higher SEs for all groups, but has an upward trend for Group B_L.

Table 4. 35. *Mean SE for the main effect of final estimate prior for both IRT models for Simulation Two, CT, peaked distributions.*

	Rasch Model	2PL Model
Population Composite Prior	0.318	0.247
Group Composite Prior	0.321	0.252
Group Specific Prior	0.323	0.255
Less Informative Prior	0.371	0.348

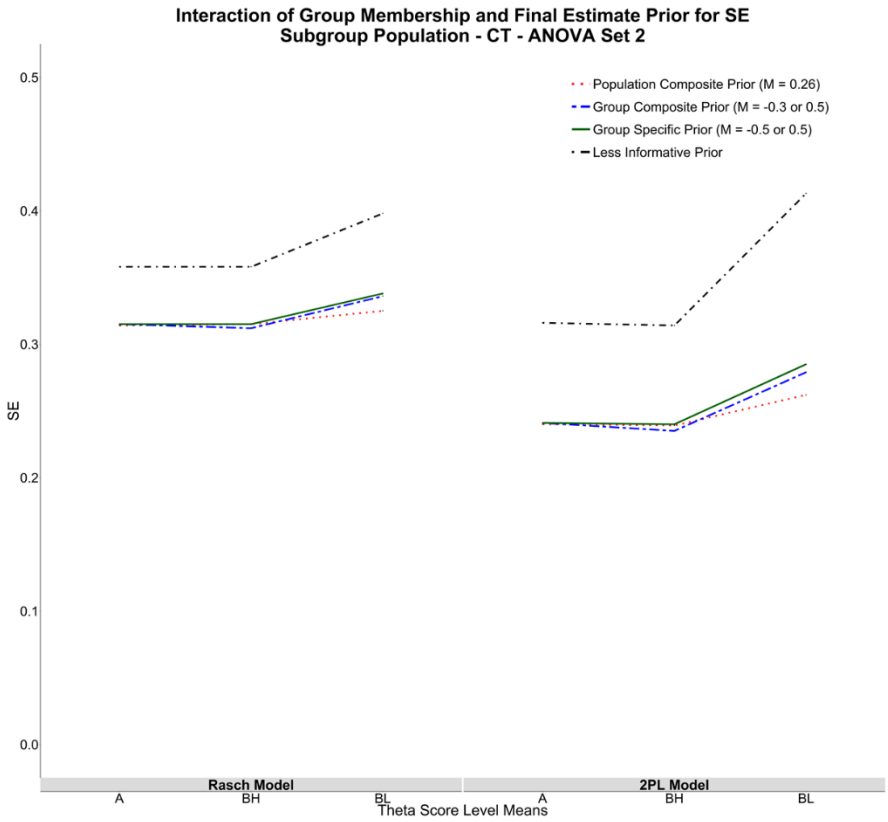

Figure 4. 28. Mean SE for the interaction between group membership and final estimate prior under both IRT models for Simulation Two, CT, peaked distributions.

Bias. There is a significant main effect for final estimate prior under both IRT models. As Table 4.36 shows, the Group Composite prior results in less bias than the other priors. The Less Informative Prior condition is comparable. The other two priors are inappropriate since their means are far from the mean of the CT. There is also an interaction between final estimate prior and group membership for the 2PL model (Figure 4.29). For the 2PL model, the three priors utilizing an informative prior underestimate ability for Groups A and B_H and overestimate ability for Group B_L. This makes sense, because the CT is easier for the higher functioning groups (A and B_H) and harder for lower groups (B_L). However, the Less Informative Prior results in lower levels of bias than the other three test priors.

Table 4. 36. *Mean bias for the main effect of final estimate prior under both IRT models for Simulation Two, CT, peaked distributions.*

	Rasch Model	2PL Model
Population Composite Prior	0.042	0.024
Group Composite Prior	0.016	0.005
Group Specific Prior	0.034	0.014
Less Informative Prior	0.026	-0.008

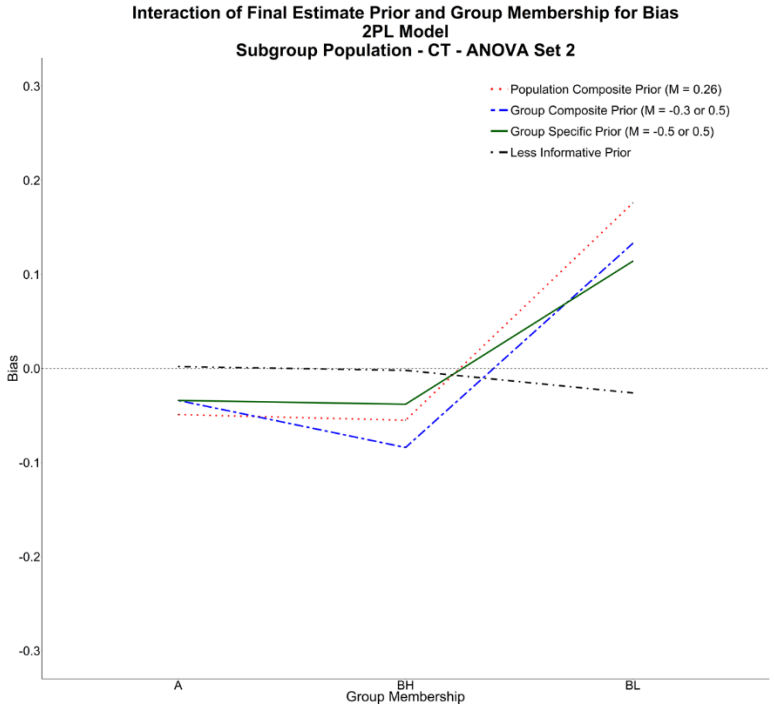

Figure 4. 29. Mean bias for the interaction between final test prior and group membership under the 2PL model for Simulation Two, CT, peaked distributions.

RMSE. There is a significant main effect for final estimate prior when examining mean RMSE (Table 4.37). For the Rasch model, the most accurate estimates are obtained using the less informative prior. However, for the 2L model, this prior produces the least accurate estimates. The Group Specific Prior is the most accurate. There is also an interaction between final estimate prior and group membership for mean RMSE for the 2PL model (Figure 4.30). The most accurate estimates are obtained using the Group Specific Prior (green). However, the

three informative test priors are similar in terms of accuracy. The less informative prior (black) produces the least accurate estimates.

Table 4. 37. *Mean RMSE for the main effect for final estimate prior under both IRT models for Simulation Two, CT, peaked distributions.*

	Rasch Model	2PL Model
Population Composite Prior	0.526	0.376
Group Composite Prior	0.514	0.367
Group Specific Prior	0.504	0.358
Less Informative Prior	0.491	0.432

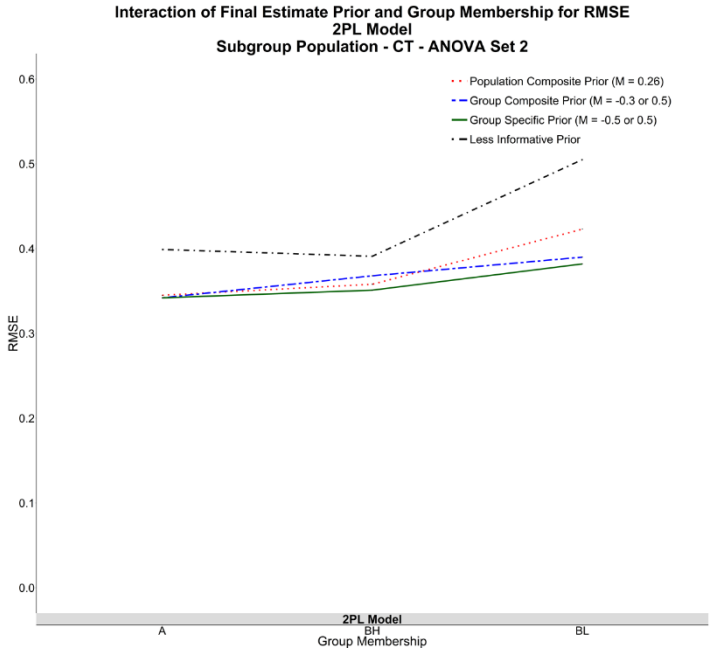

Figure 4. 30. Mean RMSE for the interaction between final estimate prior and group membership for the 2PL model for Simulation Two, CT, peaked distributions.

Other meaningful results. The remaining significant and meaningful results for Simulation Two examining the CTs for the peaked distributions is presented.

Standard error. For mean SE, there are two significant main effects – test length and group membership. Longer tests (M_R = 0.282; M_{2pl} = 0.230) produce lower mean SEs than shorter tests (M_R = 0.384; M_{2pl} = 0.321). Mean SEs are higher for Group B_L than for the other

two groups (Table 4.38). This is due to the mismatch between this group's mean ability level and the mean difficulty of the test.

Table 4. 38. *Mean SE for the main effect of group membership under both IRT models for Simulation Two, CT, peaked distributions.*

	Rasch Model	2PL Model
A	0.483	0.474
B_L	0.502	0.529
B_H	0.483	0.471

Bias. When examining bias, there is a significant main effect for test length. For CTs, the shorter tests (M_R = 0.024; M_{2pl} = 0.001) resulted in lower mean bias than longer tests (M_R = 0.035; M_{2pl} = 0.016). This might be due to more inappropriate information in terms of test items throughout the entire CT. There is also a significant main effect of group membership (Table 4.39). Group B_L has higher levels of bias; the mean of the CT (0.26) is farther away from the mean of this group than the other two groups. Group B_L abilities are generally overestimated, whereas the other two groups are underestimated. The data obtained via the CT are less informative for the lowest group (B_L), inflating the ability estimates obtained for this group. While bias levels are lower for Groups A and B_H, the CT is still less informative than an optimal test, which affects final ability estimates for the groups.

Table 4. 39. *Mean bias for the significant main effect of group membership under both IRT models for Simulation Two, CT, peaked distributions.*

	Rasch Model	2PL Model
A	-0.054	-0.029
B_L	0.225	0.099
B_H	-0.082	-0.045

RMSE. Longer tests (M_R = 0.467; M_{2pl} = 0.333) produce more accurate estimates (e.g., lower mean RMSE) than shorter tests (M_R = 0.551; M_{2pl} = 0.433). There is also a significant

main effect for group membership (Table 4.40). Estimates are more accurate for Group A, followed by Group B_H. These groups have mean abilities close to the mean of the CT. Group B_L has the least accurate ability estimates.

Table 4. 40. *Mean RMSE for the main effect of group membership under both IRT models for Simulation Two, CT, peaked distributions.*

	Rasch Model	2PL Model
A	0.484	0.357
B_L	0.547	0.425
B_H	0.495	0.367

CAT. For the CAT, test priors are utilized through the entire test administration process and for final ability estimation. Table 4.41 presents the η_w^2 for all effects and interactions for each dependent variable, separated by IRT model. Each measure meeting all previously described significance criteria will be discussed.

Table 4. 41. *Within-family effect sizes from CAT ANOVA Set 2 conducted on standard error (SE), bias, root mean square error (RMSE) for each IRT model using the Subgroup Population.*

Effect	SE		Bias		RMSE	
	Rasch	2PL	Rasch	2PL	Rasch	2PL
Results on Test Priors						
P	0.000**	0.000	**0.337****	**0.079****	0.003**	0.000
P*G	**0.071***	0.000	**0.306****	**0.096****	**0.070****	0.007
P*G*L	0.000	0.003	0.011**	0.013**	0.003	0.006
P*G*E	0.000**	0.066	**0.397****	**0.410****	**0.304****	**0.095****
P*G*E*L	0.000**	0.035	0.045**	0.012**	0.051**	0.000*
P*E	0.000**	0.000	**0.793****	**0.706****	**0.178****	0.000**
P*L	0.000	0.000	0.015**	0.005	0.000	0.000
P*E*L	0.000**	0.000	**0.087****	0.029**	0.022**	0.000**
Results Dealing with All Other Factors						
E	**0.770****	**0.935****	0.054**	**0.118****	0.022**	**0.667****
L	**0.999****	**0.937****	0.003**	0.004**	**0.933****	**0.786****
G	**0.071****	0.033**	**0.431****	**0.231****	0.012**	0.004**
E*L	**0.230****	0.065**	0.000**	0.000**	**0.200****	**0.103****
G*L	0.000**	0.001	0.025**	0.003	0.002	**0.100****
E*G	0.000**	**0.410****	**0.487****	**0.545****	0.063**	0.047**
E*G*L	**1.000****	**0.489****	0.055**	0.020**	0.013**	0.004**

* $p < 0.0167$; ** $p < 0.001$

L = test length; P = test prior

E = estimate type; G = group membership

Pertinent test prior results. As for previous sections, the results relating to the test priors are presented first.

Standard error. For mean standard error of measurement, there is only one significant and meaningful effect, the interaction between test prior and group membership for the Rasch model. However, as Table 4.42 shows, the differences in mean SE is small for all groups across all test priors; this is as expected, since the effect size barely met the criterion cut-off (0.071 vs. 0.07).

Table 4. 42. *Mean SE for the interaction between test prior and group membership under the Rasch model for Simulation Two, CAT, peaked distributions.*

	Population Composite Prior	Group Composite Prior	Group Specific Prior
A	0.280	0.280	0.280
B_L	0.282	0.280	0.280
B_H	0.280	0.282	0.279

Bias. When examining bias, there is a significant main effect of test prior for both IRT models. As shown in Table 4.43, the Group Composite Prior has the largest bias out of all the test priors and tends to underestimate ability. The Population Composite Prior has the next highest level of bias, whereas the Group Specific Prior has the lowest. For the 2PL model, the absolute bias levels are very minimal; however, the actual bias is in different directions. The effect is small (0.079) and might be driven by the polarization of bias (i.e., positive and negative).

Table 4. 43. *Mean bias for the main effect of test prior under both IRT models for Simulation Two, CAT, peaked distributions.*

	Rasch Model		2PL Model	
	Bias	Absolute Bias	Bias	Absolute Bias
Population Composite Prior	0.014	0.014	0.003	0.003
Group Composite Prior	-0.027	0.027	-0.010	0.010
Group Specific Prior	0.002	0.002	-0.001	0.001

A significant interaction between group membership and test prior exists for both IRT models, as shown in Figure 4.31. The Group Specific Prior (green) is produce the lowest levels of bias. The Population Composite Prior (red) produces bias in the minority groups (B_L and B_H). The Group Composite Prior does well for Groups A, in which the appropriate prior is being used, and Group B_L, since it is near the group's mean. However, it functions very poorly for Group B_H. It underestimates the abilities of simulees in this group.

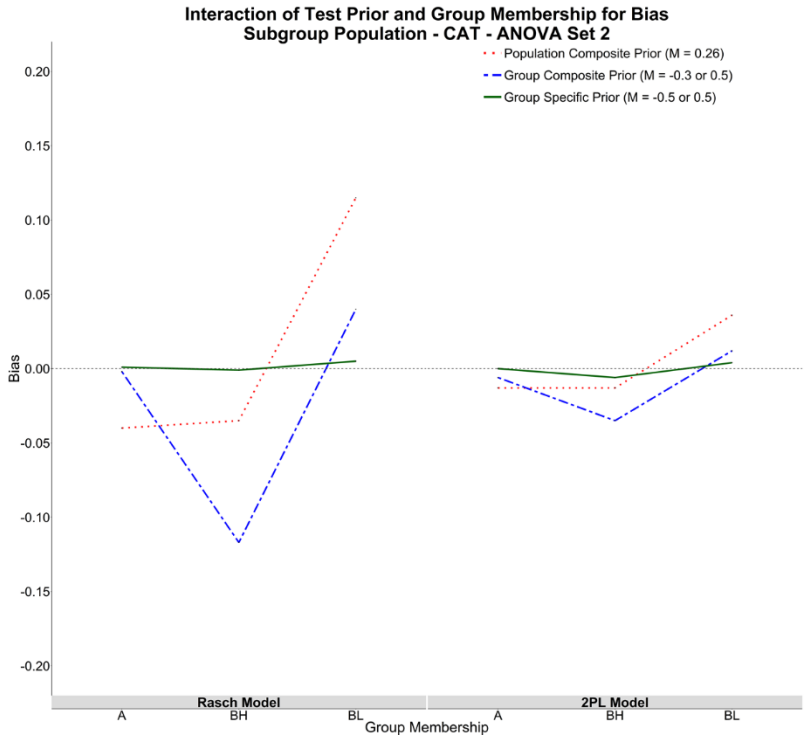

Figure 4. 31. Mean bias for the interaction between test prior and group membership for both IRT models for Simulation Two, CAT, peaked distributions.

A three-way interaction between test prior, group membership, and estimate type is significant for both IRT models when examining bias (Figure 4.32). Patterns comparable to those in Figure 4.32 above exists for the groups and test priors. However, EAP Uniform estimates do reduce the amount of bias, but do not get rid of it entirely. Not pictured is the two-way interaction using only test prior and estimate type. Patterns are comparable to those in Figure 4.31 for the three-way interaction. EAP Uniform estimates reduce the amount of bias in each of the test priors, but do not get rid of it completely.

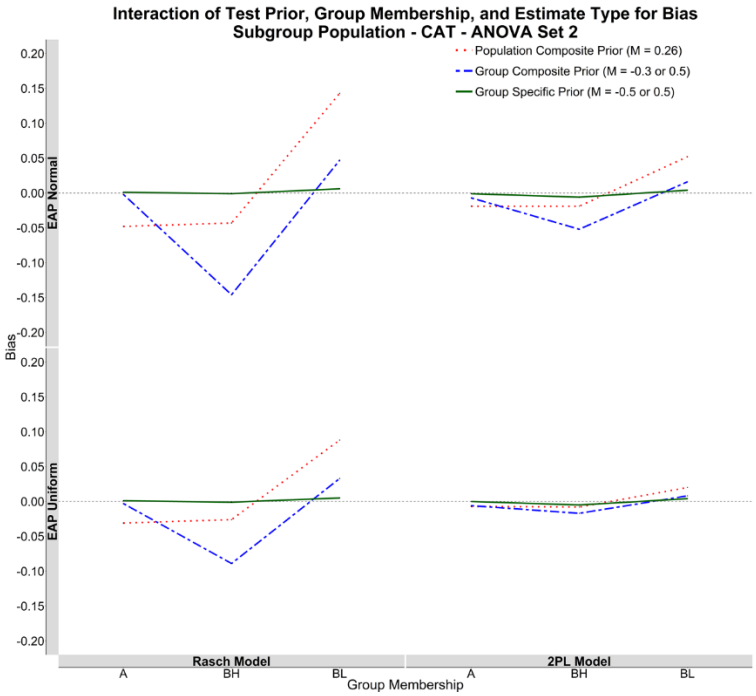

Figure 4. 32. Mean bias for the interaction between test prior, group membership, and estimate type for both IRT models for Simulation Two, CAT, peaked distributions.

The last interaction significant for the test prior in terms of bias is the three-way interaction between test prior, estimate type, and test length for the Rasch model (Figure 4.33). The Population Composite Prior (red) overestimates ability whereas the Group Composite Prior (blue) underestimates ability. The Group Specific Prior (green) has very little bias. However, a reduction in bias is observed for the Group Composite Prior (blue) when a longer test is utilized.

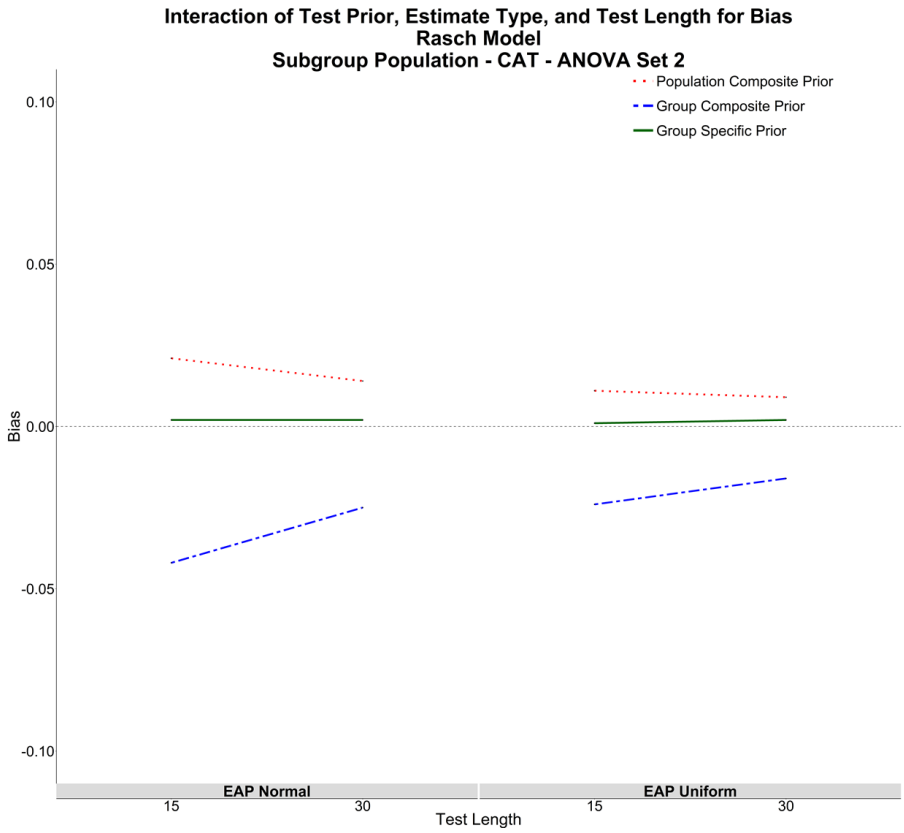

Figure 4. 33. Interaction between test prior, estimate type, and test length for mean bias under the Rasch model for Simulation Two, CAT, peaked distributions.

RMSE. Concerning RMSE, the interaction between test prior, group membership, and estimate type is significant for both IRT models, and is presented in Figure 4.34. For Group A (A), all three priors are similar. For Group B_L (B), the Population Composite Prior (red) is not as accurate as the group priors. These priors function similarly in terms of accuracy because their means are close. For Group B_H (C), the Population Composite Prior (red) and Group Specific Prior (green) function similar. The Group Composite is less accurate for these simulees. There is a significant two-way interaction for RMSE under the Rasch model between test prior and estimate type, in which the Subgroup Prior using EAP Normal as a final estimate approach produces the most accurate estimates. For this test prior, EAP Uniform is also lower for all other test priors.

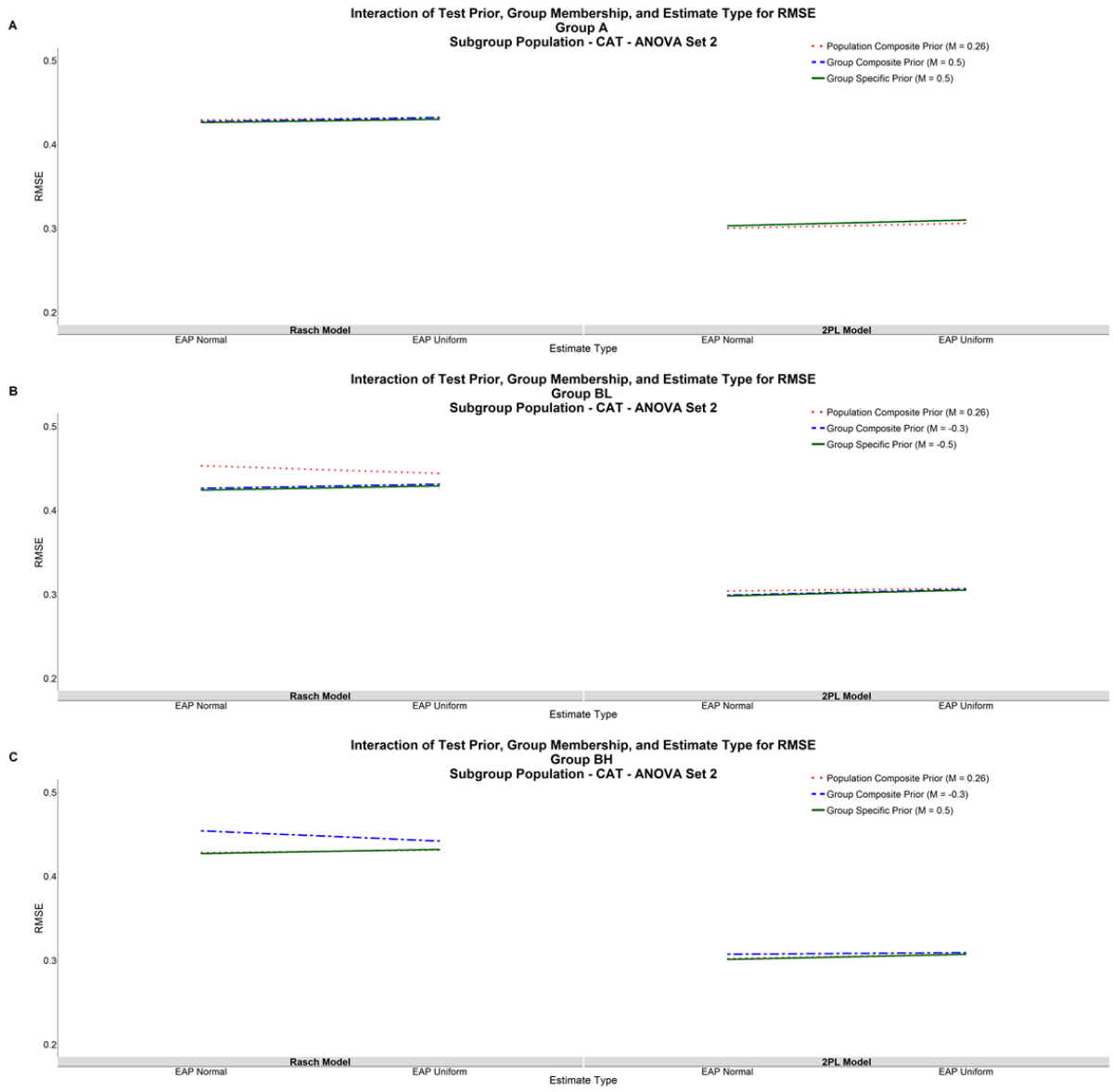

Figure 4. 34. Interaction of test prior, group membership, and estimate type for mean RMSE for both IRT models for Simulation Two, CAT, peaked distributions.

Other meaningful results. The remaining results concerning the CATs for Simulation Two using the peaked distributions is presented.

Standard error. For both IRT models, EAP Normal (M_R = 0.272; M_{2pl} = 0.211) produces lower mean SEs than EAP Uniform (M_R = 0.288; M_{2pl} = 0.216). Longer tests (M_R = 0.231; M_{2pl} = 0.195) also produce lower mean SEs than shorter tests (M_R = 0.330; M_{2pl} = 0.232) for both IRT models. For SE under the Rasch model, there is also a significant interaction for estimate type

and test length. Table 4.44 shows that tests using EAP Uniform have higher mean SEs, but the longer the test, the closer the two estimation methods are in terms of SE.

Table 4. 44. *Mean SE for the interaction between estimate type and test length for the Rasch model for Simulation Two, CAT, peaked distributions.*

	EAP Normal	EAP Uniform
15	0.318	0.342
30	0.227	0.234

There is a significant interaction between estimate type and group membership for SE under the 2PL model. In Table 4.45, the EAP Normal approach produces lower mean SEs for all three groups. Finally, for SE under both IRT models, there is a significant interaction between estimate type, group membership, and test length (Figure 4.35). For all groups, EAP Normal produces the lowest mean SEs, but the two estimate types are approximately equal for a longer test.

Table 4. 45. *Mean SE for the interaction between estimate type and group membership for the 2PL model for Simulation Two, CAT, peaked distributions.*

	EAP Normal	EAP Uniform
A	0.211	0.217
B_L	0.210	0.215
B_H	0.211	0.217

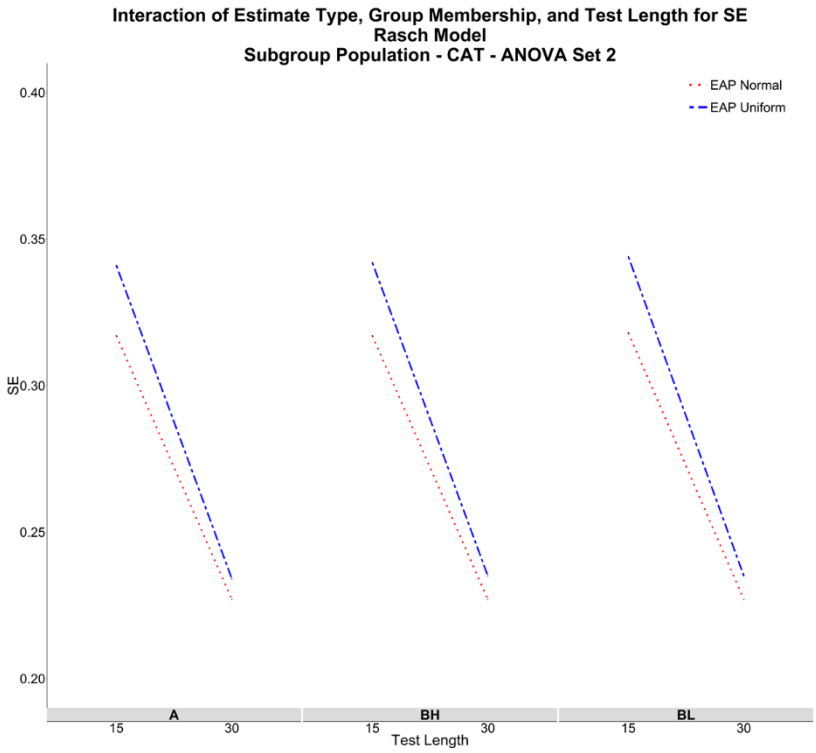

Figure 4. 35. Mean SE for the interaction between estimate type, group membership, and test length under the Rasch model for Simulation Two, CAT, peaked distributions.

Bias. For both IRT models, a significant main effect for estimate type exists for mean bias for the 2PL model. EAP Uniform (M_{2pl} = -0.001) produces less bias than EAP Normal (M_{2pl} = -0.004), but the differences are minimal. There is also a significant main effect for group membership for both IRT models. Table 4.46 presents the bias, as well as absolute bias, for the three groups. Group A tends to have less bias than the other groups; more of the test priors are appropriate for this group. Groups B_L and B_H have approximately the same levels of mean bias.

Table 4. 46. *Mean bias for the main effect for group membership under both IRT models for Simulation Two, CAT, peaked distributions.*

	Rasch Model		2PL Model	
	Bias	Absolute Bias	Bias	Absolute Bias
A	-0.014	0.014	-0.007	0.007
B_L	0.054	0.054	0.017	0.017
B_H	-0.051	0.051	-0.018	0.018

There is a significant interaction between estimate type and group membership for both IRT models, as well as a three-way interaction between estimate type, group membership, and test length (Figure 4.36). EAP Uniform estimates have less absolute bias when compared to EAP Normal for all three groups. This pattern is noticeable for the minority groups (B_L and B_H) who may have received more inappropriate priors.

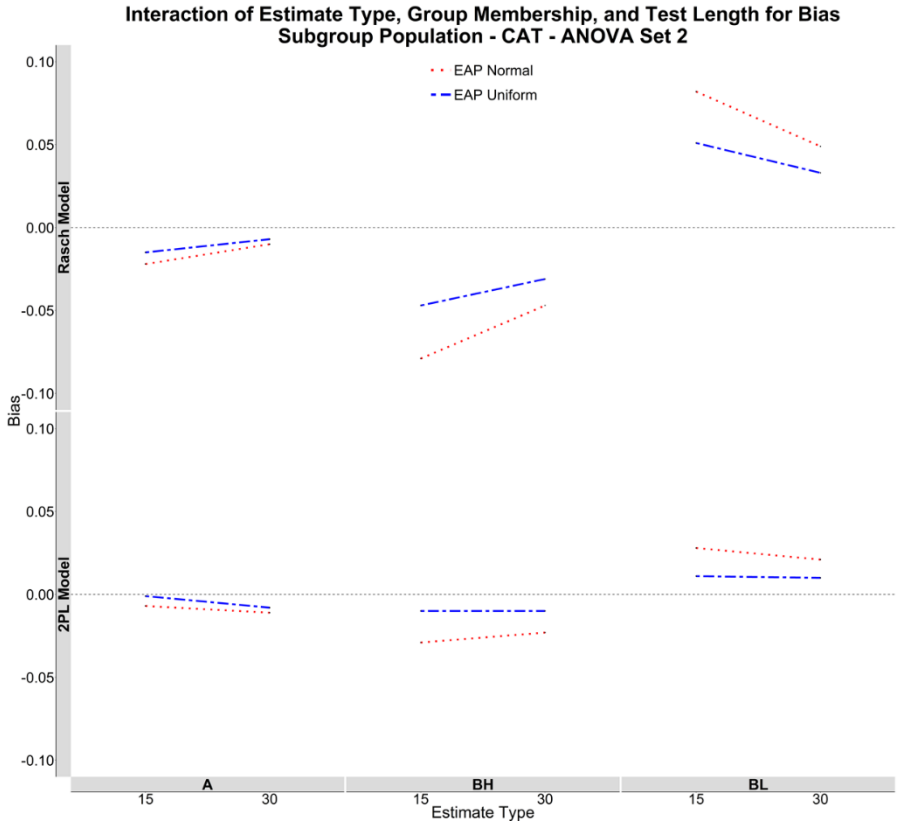

Figure 4. 36. Mean bias for the interaction between estimate type, group membership, and test length under both IRT models for Simulation Two, CAT, peaked distributions.

RMSE. For the 2PL model, there is a significant main effect for estimate type, in which EAP Normal estimates (M = 0.302) produce slightly more accurate estimates than EAP Uniform (M = 0.307). However, the means are still similar. There is also a significant main effect for test length. Longer tests (M_R = 0.367; M_{2pl} = 0.277) produce more accurate estimates than shorter tests (M_R = 0.499; M_{2pl} = 0.332). There is also a significant interaction between estimate type

and test length for both IRT models (Table 4.47). EAP Normal estimates are a little more accurate, but the two estimate types are approximately the same, especially for longer tests. Lastly, there is an interaction between group membership and test length. Although not explicitly stated, the longer test for all groups produces more accurate estimates, as expected, since the main effect of test length was significant.

Table 4. 47. *Mean RMSE for the interaction between test length and estimate type under both IRT models for Simulation Two, CAT, peaked distributions.*

	Rasch Model		2PL Model	
	EAP Normal	EAP Uniform	EAP Normal	EAP Uniform
15	0.497	0.501	0.328	0.336
30	0.368	0.366	0.275	0.279

ANOVA Set 3. The third set of ANOVAs for Simulation Two represent a set of data in which the population composition is fully intact (i.e., Group B_H composes only 6% of the data and all separate distributions exist); group membership is examined in this set of ANOVAs. This approach allowed for the examination of how the test priors would function in the true population, and the interaction with group membership. As with all previous sections, CT results are first presented followed by CAT results.

CT. Similar to ANOVA Set 2 for the CTs, ANOVA Set 3 examined the groups and their interactions with the final estimate priors. However, instead of each group having an equal number of simulees, Group B_H only had 600 simulees (i.e., 6% of the population). As with other CT conditions, the priors only influence final estimates. Table 4.48 presents the effect size information for all effects and interactions.

Table 4. 48. *Within-family effect sizes from CT ANOVA Set 3 conducted on standard error (SE), bias, root mean square error (RMSE) for each IRT model using the Subgroup Population.*

Effect	SE		Bias		RMSE	
	Rasch	**2PL**	**Rasch**	**2PL**	**Rasch**	**2PL**
Results on Final Estimate Priors						
P	**0.136****	**0.350****	0.013	0.048**	0.029**	**0.176****
P*G	**0.101****	**0.150****	0.038**	**0.220****	0.012**	0.024**
P*G*L	0.015**	0.006**	0.004**	0.022**	0.002	0.005
P*L	0.034**	0.027**	0.005	0.010	0.012**	0.042**
Results Dealing with All Other Factors						
L	**0.805****	**0.548****	0.006	0.029**	**0.444****	**0.463****
G	**0.551****	**0.562****	**0.775****	**0.404****	**0.366****	**0.395****
G*L	0.019**	0.007**	0.000	0.001	0.005*	0.002

* $p < 0.0167$; ** $p < 0.001$
P = final estimate prior; L = test length
G = group membership

Pertinent test prior results. All results relating to the final estimate priors are presented before presentation of all other results.

Standard error. A significant main effect for final estimate prior exists when examining mean SE (Table 4.49). For the informative final estimate priors, all levels of mean SE are relatively similar. However, for the Less Informative Prior, mean SE is higher. There is also a significant interaction between final estimate prior and group membership (Figure 4.37). The Less Informative Prior condition has a higher SE mean for all three groups, but has an upward trend for Group B_L.

Table 4. 49. *Mean SE for the main effect of final estimate prior for both IRT models for Simulation Two, CT, population distributions.*

	Rasch Model	2PL Model
Population Composite Prior	0.314	0.237
Group Composite Prior	0.319	0.245
Group Specific Prior	0.320	0.249
Less Informative Prior	0.365	0.331

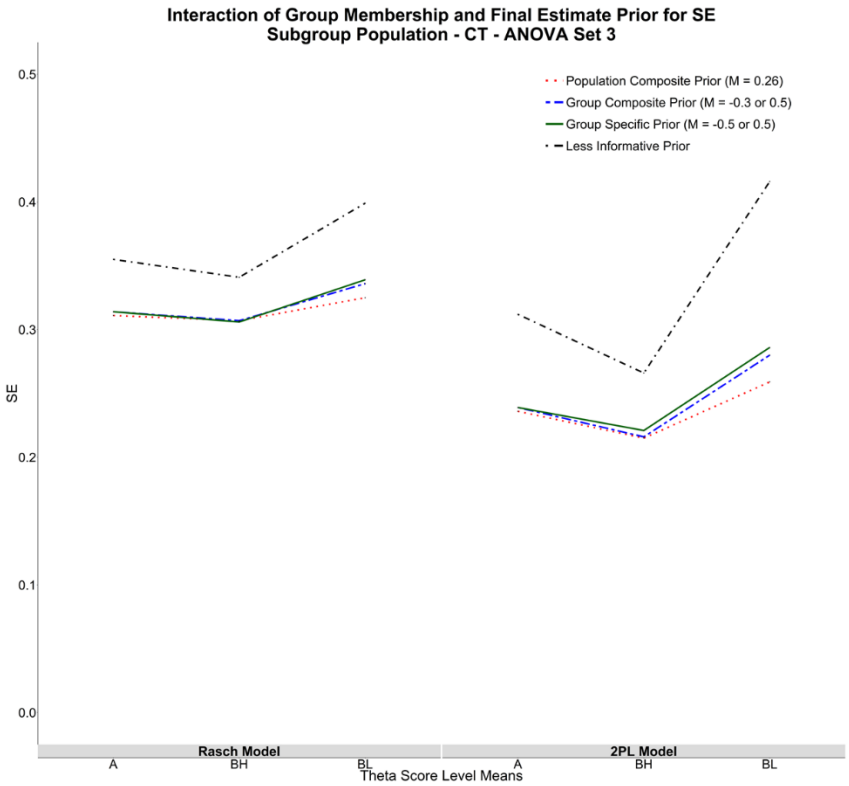

Figure 4. 37. Mean SE for the interaction between group membership and final estimate prior under both IRT models for Simulation Two, CT, population distributions.

Bias. In terms of bias, the only significant interaction as between the final estimate prior and group membership (Figure 4.38) for the 2PL model. As shown in the figure, the three final estimate priors utilizing an informative prior underestimate ability for Groups A and B_H, but overestimate ability for Group B_L. The Less Informative Prior, however, has less of an influence on bias for all three groups.

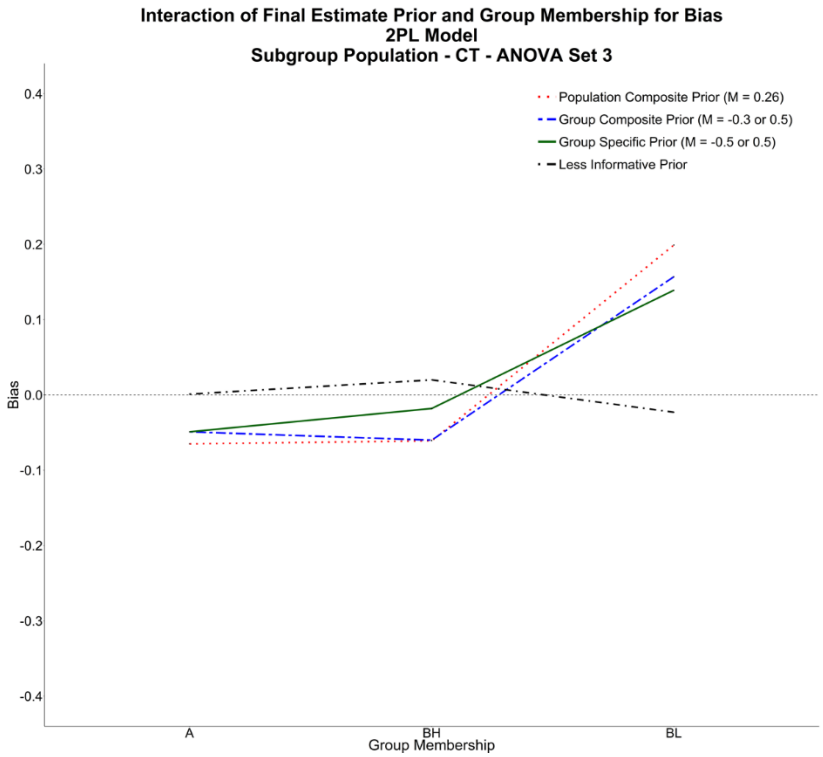

Figure 4. 38. Mean bias for the interaction between final estimate prior and group membership under the 2PL model for Simulation Two, CT, population distributions.

RMSE. For RMSE, the main effect of final estimate prior for the 2PL model is significant. The Group Specific Prior (M = 0.332) produces the most accurate estimates. The Group Composite Prior (M = 0.339) is almost as accurate, and is more accurate than the Population Composite Prior (M = 0.350). The Less Informative Prior (M = 0.416) produces the least accurate estimates.

Other meaningful results. The remaining significant and meaningful results for ANOVA Set 3 concerning the CTs are presented.

Standard error. Longer tests (M_R = 0.280; M_{2pl} = 0.218) result in lower levels of mean SE than shorter tests (M_R = 0.379; M_{2pl} = 0.313). There is also a significant main effect for group membership (Table 4.50) for both IRT models. Mean SEs are higher for group B_L, which is

primarily due to the ability-difficulty mismatch between the true ability of the group and the difficulty of the CT.

Table 4. 50. *Mean SE for the main effect of group membership under both IRT models for Simulation Two, CT, population distributions.*

	Rasch Model	2PL Model
A	0.323	0.256
B_L	0.350	0.311
B_H	0.315	0.230

Bias. When examining mean bias, there is a significant main effect for group membership. Table 4.51 presents the information for each group under each IRT model. Groups A and B_H are underestimated; the mean of the CT is under the ability of these groups, suggesting not enough items are present at the appropriate level to obtain best estimates. However, bias levels are low. For Group B_L, the abilities are overestimated. This is because the items are above the mean of the group and are harder for the simulees.

Table 4. 51. *Mean bias for the main effect of group membership under both IRT models for Simulation Two, CT, population distributions.*

	Rasch Model	2PL Model
A	-0.077	-0.041
B_L	0.258	0.118
B_H	-0.066	-0.030

RMSE. When examining mean RMSE, two main effects are significant – test length and group membership. Longer tests (M_R = 0.442; M_{2pl} = 0.305) result in more accurate estimates than shorter tests (M_R = 0.527; M_{2pl} = 0.413). For group membership (Table 4.52), Groups A and B_H receive more accurate estimates than Group B_L.

Table 4. 52. *Mean RMSE for the main effect of group membership under both IRT models for Simulation Two, CT, population distributions.*

	Rasch Model	2PL Model
A	0.486	0.353
B_L	0.545	0.424
B_H	0.423	0.300

CAT. ANOVAs were conducted using Sample Set 3 for the CATs. These assessments utilize the test priors throughout the entire test process. Table 4.53 presents the within-family effects for each dependent variable for all variables. Less effects/interactions were flagged as meaningful when using the population distributions.

Table 4. 53. *Within-family effect sizes from CAT ANOVA Set 3 conducted on standard error (SE), bias, root mean square error (RMSE) for each IRT model using the Subgroup Population.*

Effect	**SE**		**Bias**		**RMSE**	
	Rasch	**2PL**	**Rasch**	**2PL**	**Rasch**	**2PL**
Results on Test Priors						
P	0.000*	0.000	**0.093****	0.020*	0.002	0.002
P*G	0.034**	0.000	**0.121****	0.020**	0.022**	0.004
P*G*L	0.000	0.001	0.013**	0.007	0.003	0.001
P*G*E	0.000**	0.024	**0.290****	**0.319****	**0.108****	0.019**
P*G*E*L	0.000**	0.017	0.029**	0.007**	0.025**	0.009
P*E	0.000**	0.000	**0.587****	**0.639****	0.040**	0.010**
P*L	0.000	0.000	0.014*	0.006	0.003	0.000
P*E*L	0.000**	0.000	0.055**	0.000**	0.000	0.001
Results Dealing with All Other Factors						
E	**0.769****	**0.931****	0.009**	0.056**	**0.159****	**0.450****
L	**0.995****	**0.902****	0.000	0.001	**0.0667*****	**0.366****
G	0.057**	0.031**	**0.208****	0.054**	0.004	0.028**
E*L	**0.227****	0.069**	0.001	0.001	**0.139****	**0.070****
G*L	0.000	0.000	0.010**	0.002	0.001	0.003
E*G	0.000**	**0.760****	**0.537****	**0.603****	0.028**	0.009**
E*G*L	0.024**	**0.199****	0.063**	0.023**	0.003	0.000

* $p < 0.0167$; ** $p < 0.001$

L = test length; P = test prior

E = estimate type; G = group membership

Pertinent test prior results. First, all results relating to the three test priors are presented for each of the dependent measures of interest. It should be noted that no significant effects exist for standard error.

Bias. A significant main effect for test prior is found for the Rasch model (Table 4.54). The Group Specific Prior has the lowest level of bias, followed by the Population Composite Prior. The Group Composite Prior has the highest absolute level of bias, and underestimates abilities.

Table 4. 54. *Mean bias for the main effect of test prior under the Rasch model for Simulation Two, CAT, population distributions.*

	Bias	Absolute Bias
Population Composite Prior	0.019	0.019
Group Composite Prior	-0.029	0.029
Group Specific Prior	0.005	0.005

For the Rasch model, there is also a significant interaction between group membership and test prior. A significant two-way interaction between test prior and estimate type exists for both IRT models as well. However, there is also a significant interaction between test prior, group membership, and estimate type for both IRT models, and is presented in Figure 4.39. As the figure shows, the Group Specific Prior (green) has less of an influence in terms of bias for all three groups. For Group A, the two group priors function the same; this is expected, since the mean of these priors are the same for this group. For Group B_L, the Group Composite Prior (blue) does overestimate abilities, but the Population Composite Prior (red) overestimates them even more. This is due to the increased influence of the Population Composite Prior on the estimates for this group. For Group B_H, the Group Composite Prior (blue) underestimates abilities; it has more of an influence of final ability estimation than the other three priors.

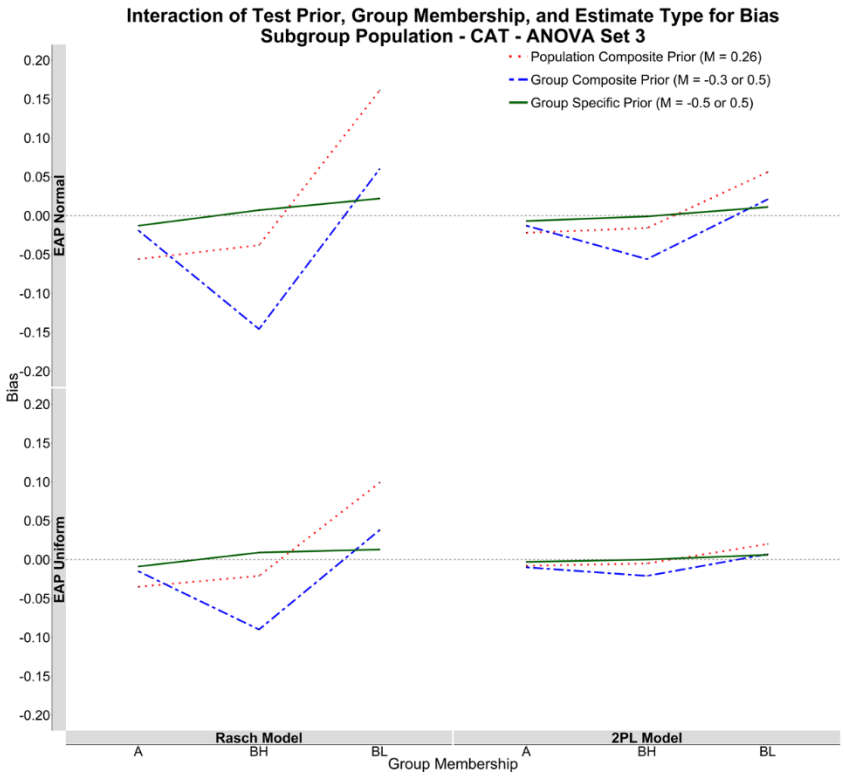

Figure 4. 39. Mean bias for the interaction between test prior, group membership, and estimate type for both IRT models for Simulation Two, CAT, population distributions.

RMSE. For mean RMSE values, only the interaction between test prior, group membership, and estimate type is significant for the Rasch model (Figure 4.40). For Group A (Figure 4.40A), the test priors all provide relatively similar accuracy levels. For Group B_L (Figure 4.40B), the Population Composite Prior (red) provides the least accurate estimates, whereas the two group priors are similar. For Group B_H (Figure 4.40C), the Group Composite Prior (blue) provides the lowest levels of accuracy. While the Population Composite Prior (red) is slightly lower than the Group Specific Prior (green), the differences are minimal.

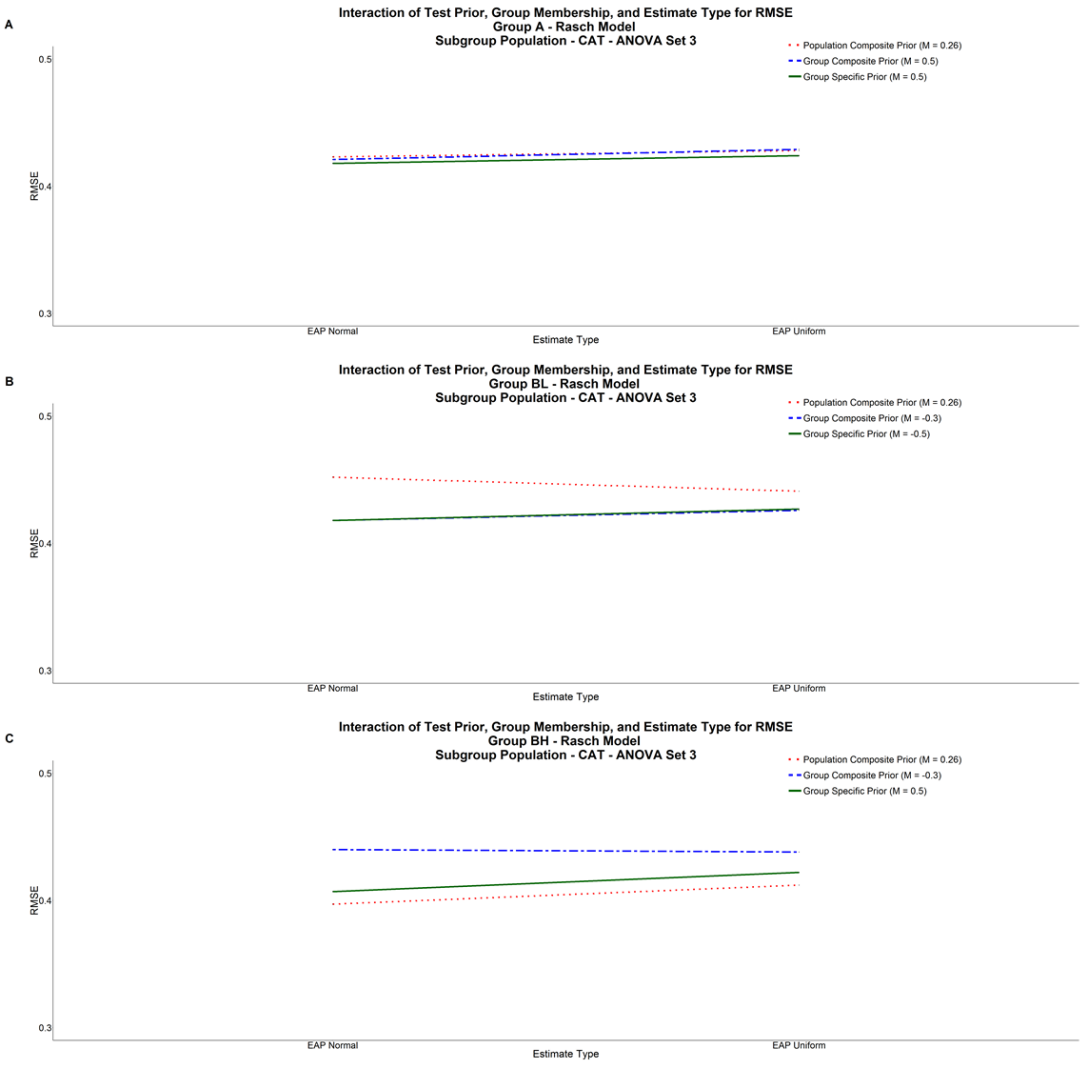

Figure 4. 40. Mean RMSE for the interaction between test prior, group membership, and estimate type for the Rasch model for Simulation Two, CAT, population distributions.

Other meaningful results. The remaining meaningful results for the CAT population distributions for Simulation Two, Sample Set 3, are presented.

Standard error. The main effects for estimate type and test length are significant for both IRT models. The EAP Normal (M_R = 0.272; M_{2pl} = 0.210) estimate approach to final trait estimation produces lower mean SE than the EAP Uniform (M_R = 0.287; M_{2pl} = 0.216) estimate approach. Longer tests (M_R = 0.230; M_{2pl} = 0.194) also produce lower levels of mean SE than shorter tests (M_R = 0.329; M_{2pl} = 0.232). For the Rasch model, the interaction between test

length and estimate type is significant (Table 4.55). Longer tests using EAP Normal for final trait estimation result in lower mean SEs.

Table 4. 55. *Mean SE for the interaction between estimate type and test length for the Rasch model for Simulation Two, CAT, population distributions.*

	EAP Normal	EAP Uniform
15	0.317	0.340
30	0.226	0.233

There is a significant interaction for the 2PL model between estimate type and group membership (Table 4.56). The EAP Normal approach produces lower mean SEs than the EAP Uniform approach. Not presented is the significant interaction between estimate type, group membership, and test length for the 2PL model; however, the results are like those in Figure 4.34. Longer tests, in conjunction with EAP Normal estimates, produce lower mean SEs for all groups.

Table 4. 56. *Mean SE for the interaction between estimate type and group membership for the 2PL model for Simulation Two, CAT, population distributions.*

	EAP Normal	EAP Uniform
A	0.211	0.217
B_L	0.210	0.216
B_H	0.209	0.215

Bias. There is a significant main effect for the Rasch model when examining mean bias for the three groups. As Table 4.57 shows, Group B_L has the highest levels of mean bias, followed by Group B_H. Group A has the lowest levels of bias. The abilities for the two high-functioning groups (A and B_H) are underestimated, while the abilities for Group B_L are overestimated. There is a significant interaction between estimate type and group membership for both IRT models (Table 4.58). The EAP Uniform final ability estimate approach has less of an influence over the estimates, resulting in lower levels of bias for both IRT models than the

EAP Normal final ability estimate. As with the main effect of group membership, abilities are underestimate for Groups A and B_H overestimate for Group B_L.

Table 4. 57. *Mean bias for the main effect of group membership under the Rasch model for Simulation Two, CAT, population distributions.*

	Bias	Absolute Bias
A	-0.024	0.024
B_L	0.065	0.065
B_H	-0.046	0.046

Table 4. 58. *Mean bias for the interaction between estimate type and group membership under both IRT models for Simulation Two, CAT, population distributions.*

	Rasch Model		2PL Model	
	EAP Normal	EAP Uniform	EAP Normal	EAP Uniform
A	-0.029	-0.019	-0.014	-0.007
B_L	0.081	0.050	0.029	0.011
B_H	-0.059	-0.034	-0.024	-0.009

RMSE. The main effects of estimate type and test length are both significant for mean RMSE. The EAP Normal (M_R = 0.422; M_{2pl} = 0.295) resulted is slightly more accurate ability estimates than EAP Uniform (M_R = 0.428; M_{2pl} = 0.302). As with all other analyses, longer tests (M_R = 0.359; M_{2pl} = 0.273) result in more accurate estimates than shorter tests (M_R = 0.490; M_{2pl} = 0.324). There is also a significant interaction between estimate type and test length for both IRT models (Table 4.59). As expected, longer assessments utilizing the EAP Normal estimate produce the most accurate final ability estimates.

Table 4. 59. *Mean RMSE for the interaction between estimate type and test length under both IRT models for Simulation Two, CAT, population distributions.*

	Rasch Model		2PL Model	
	EAP Normal	EAP Uniform	EAP Normal	EAP Uniform
15	0.484	0.496	0.319	0.329
30	0.359	0.360	0.270	0.275

CHAPTER 5

DISCUSSION

The final chapter includes a discussion of the major findings of the study, as well as their implications. While various findings are highlighted, focus is placed on the relationship of the test prior utilized during the administration of the assessment, primarily the CAT, in conjunction with the final estimate approach (i.e., EAP Normal vs. EAP Uniform) used to obtain the final ability estimate. Limitations of the study are also presented. The paper concludes with recommendations for further research as elucidated by the current study.

Discussion of Findings

A multitude of effects were examined in this simulation study. Highlights of various findings are presented, but primary focus is on interaction between the use of the different test priors during test administration and the final ability estimate method, in conjunction with various person and ability groups, for final trait estimation. Globally, many results were expected. For example, in both simulations, CATs perform better than CTs (conventional tests) in terms of theta recovery, standard error, bias, and RMSE. Estimates obtained using a tailored test more closely approximated true ability than those obtained using a static test. This result was expected, since the ability-difficulty mismatch common in CTs is resolved by using a CAT. This result also can be seen when comparing the global results from Simulation Two between the CTs and CATs. Another expected finding was that, in general, longer assessments perform better than shorter assessments, especially for a CAT. Further, when examining abilities at various locations along the latent trait continuum, individuals at the extreme ends of the continuum were less accurately estimated than those towards the middle. This occurs for extreme abilities because, often, information obtained from the test prior used for ability

estimation is more discrepant. Also, the data are often less informative, especially for the CTs, because fewer appropriate items exist, which influences estimation for extreme abilities. Lastly, as expected, estimates resulting from EAP with a designated prior (i.e., EAP Normal) do regress the estimates to the mean of the prior. Often, this regression has the largest impact at the extremes of the continuum, with the prior used has the most influence. Therefore, there is less accuracy associated with this estimation approach than when using a less informative prior (i.e., EAP Uniform). However, there is a trade-off, because although EAP Uniform estimates have less bias, they often result in larger SEs overall.

The primary focus of the study was the utilization of different test priors, some of which were more appropriate than others, for item selection and final ability estimation in a tailored assessment (CAT). Two approaches to ability estimation were examined: an informative test prior was used to initialize the CAT and to select items within the assessment, and then this informative prior used during testing was either used to obtain the final ability estimate or a less informative prior was used. Thus, if the test prior used during administration is inappropriate, less informative data (i.e., the scored responses) may be obtained since the items selected may not be the most appropriate for the individual. Using a less informative prior for final trait estimation may aid in trait estimation when information from the data is scarce, but the prior might have a larger influence than anticipated over the ability estimate. Influence of the test prior was not only important on a macro level (e.g., how did the test prior function overall), but also on a micro level (e.g., how did the test prior function with the various subgroups). The test priors examined were based off the existence of different subgroups in the population who had different distributions in terms of mean ability. Therefore, how the test priors influenced the estimates for these groups was of utmost importance (i.e., Simulation Two, Sample Sets 2 and 3).

In terms of standard error, test priors do have an influence, but often these influences are rather small. To ensure lower SEs, using a longer test seems to have the most influence, as all test priors were relatively equivalent. These results were as expected. Lastly, using EAP Normal for final trait estimation leads to slightly lower mean SEs for each of the priors. Test priors have interactions with group membership in terms of bias and RMSE (i.e., Sample Sets 2 and 3 involving Simulation Two). When the population composition is examined (Sample Set 3), less effects are significant, but similar patterns are observed. Thus, an investigation into how the prior function across the subgroups when the population mean abilities are represented is warranted.

In terms of overall bias for the test priors for Simulation Two, Sample Sets 2 and 3, the Group Composite Prior had the largest bias and tends to underestimate ability. While the Group Composite Prior is appropriate for Group A, it is inappropriate for Groups B_H and B_L. The mean ability of these groups (0.5 and -0.5, respectively) is discrepant from the mean of the prior for Group B as a while (-0.3). While the mean for Group B_L is more congruent with the prior mean, Group B_H has a mean ability that is much higher. Therefore, an investigation into the interaction between test prior and group membership answered why the test priors are functioning differently at a macro-level for bias, as summarized below.

For Group A, or the majority group, the two group-based priors (i.e., Group Composite and Group Specific) are the same (Figure 4.32). That is, as the two priors utilize the same mean in the prior (M = 0.5). The Population Composite Prior, in contrast, produced more bias, and thus less accurate estimates. Ability is underestimated when using this prior. These results are expected, as the mean of the prior (M = 0.26) is below the mean of the group (0.5), so some underestimation is expected due to regression to the mean. Generally, EAP Uniform estimates

had less bias, but the difference was nonexistent when the appropriate priors were used (i.e., group priors) and similar for the Population Composite Prior.

For Group B_L, which is the low-ability subgroup of the minority group, the two group-based priors provide more accurate estimates. The Group Specific Prior (M = -0.5) provided less biased, and more accurate estimates. This is the most appropriate prior for the group, since the test prior mean matched the mean of the group (M = -0.5). The next best test prior was the Group Composite Prior (M = -0.3). This prior functioned better than the Population Composite Prior (M = 0.26). While the Group Composite Prior is not the most appropriate prior for this group, it is more appropriate than the Population Composite Prior. The Population Composite Prior tends to overestimate ability; this is expected, since the mean of the prior is higher than the mean of the group and lower ability levels are generally overestimated (as seen through examination of the theta score levels). Using EAP Uniform estimates results in similar levels of bias and accuracy for the three priors, but the pattern described still holds.

Lastly, the impact of the various test priors on Group B_H was investigated. While this subgroup is part of the minority, it is a higher-ability group that has a distribution similar to the majority group (A). Thus, the test priors might differentially affect this group than they did with the other two groups. The Group Specific Prior (0.5) produces the most accurate estimates, as it is the most appropriate prior. However, unlike the other two groups, the Population Composite Prior provides a better estimate than the Group Composite Prior for Group B_H. In this case, the Group Composite Prior (M = -0.3) is the least appropriate test prior because its mean is the farthest from that of the group (M = 0.5). The Population Composite Prior (M = 0.26), while still not the most appropriate, is closer. For this group, the test priors tend underestimate ability, except for Group Specific, but the Group Composite Prior leads to more negative bias. This

group is being adversely affected using a prior based on the composite mean of the individual group rather than either a specific group prior, which is unobtainable, or a population prior.

The previous results for the CATs were compared to those concerning the CTs under Simulation Two, Sample Set 2, where multiple subgroups existed with various means. The Group Specific Prior for the CT results in lower bias than any of the other final estimate priors based on group membership for all three groups. For Group A, the Group Specific Prior and the Group Composite Prior functioned similarly. The Group Composite Prior for final trait estimation underestimates ability for Group B_H, but overestimates ability for Group B_L. These results must consider the difficulty of the CT. The mean of the CT is 0.26, which is discrepant from the mean of the Group Composite Prior for both subgroups in B. While these two group-based final estimate priors had more bias, they did provide more accurate estimates. The Less Informative Prior condition led to less bias but higher RMSEs in the 2PL model; it had less information pertaining to the simulee and thus relied on informative data in terms of the scored responses. Using the scored responses and the item parameters, inaccurate information was provided for the simulees. While the Less Informative Prior final estimate approach might result in less bias, it had more of an influence on SEs (e.g., higher) provided less accurate information; this is different from the results found with the CAT, which showed that a less informative prior was often more accurate. Therefore, informative priors are better for CTs. The Group Composite Prior, as in the CATs, does adversely impact the Group B subgroups, and thus, is the most inappropriate for the same reasons as above. That is, the means of the Group Specific Prior are more aligned to the subgroup means, but are unattainable. The Group Composite Prior and Population Composite Priors do adversely affect subgroups of varying ability levels, but to different degrees.

Implications of Findings

A concern of the paper was whether the use of group specific priors would be beneficial to use during testing for various subgroups. One major concern was whether the use of these group composite priors would be appealing to a testing company, not just in terms of improved psychometric qualities (e.g., improved measurement), but also for other reasons. For example, if the Group Composite Priors provide improved measurement over Population Composite Priors, but they introduce concerns among the test population, the testing company may be more susceptible to legal issues because of the adverse impact of the Group Composite Priors on groups with varying levels of ability. Thus, overall, the implementation of these priors may not be beneficial for high-stakes testing (i.e., summative assessments). However, they may prove useful in recurring formative assessments, such as those given in classrooms.

Based on the findings of the study, the concluding recommendation is to continue using population-based priors for test administration and ability estimation, even if group-based priors are known. The level of bias and accuracy of estimates across the ability continuum do vary for the groups based on the test prior used. Even if group specific test priors result in the most accurate ability estimates overall, true ability is unknown and therefore, individuals cannot be placed into the appropriate subgroup (e.g., high or low ability) to have any gain in ability estimation. The high functioning minority group is adversely affected when a low, inappropriate prior is used to estimate abilities. While the impact might be minimal from a psychometric standpoint, legal and political implications outweigh these potential benefits.

Recommendations, aside from test priors, can also be given. Regarding test length, 30-items seems to be an appropriate number of items. A goal of testing is to provide enough items to obtain an accurate estimate of ability without bombarding the examinee with too many items

to affect their performance (e.g., fatigue). Thus, a test with 30 items appears to be sufficient. If possible, adaptive tests provide improved measurement over static tests. If available, these tests should be utilized. However, other issues, such as cost, item bank issues, and others, might prevent their use. Therefore, CTs are often used. The less informative final estimate prior has high SEs but less bias, so it is possible to use only the scored responses. However, if SE is a concern, the Population Composite is a viable option. Lastly, recommendations can be given in terms of final trait estimation. If standard errors are the primary concern, a test should utilize a Bayesian approach with an informed prior (e.g., EAP Normal). This prior generally results in lower standard errors. However, if bias/accuracy is the primary concern, the test administrator may choose to use a Bayesian approach with an uninformed prior (e.g., EAP Uniform). In this current study, this approach resulted in more accurate estimates across the entire ability continuum.

Limitations

As with all simulation studies, a major limitation of this research is that the data are simulated and not real. When data are simulated, a high degree of control is placed on both the data and the test design. Unfortunately, real test data are often even messier and therefore might result in different results. These results might be caused by the item pools; in this study, while the item pools aimed to simulate real-world scenarios, there was still an elevated level of control that may be absent in real item banks/pools. Or, the results could be caused by the test examinees. Often, responses by examinees are affected by different things, such as the external environment (e.g., room temperature, subliminal noises) or the person's internal environment (e.g., fatigue, hunger), and may not always result in expected patterns of responding. These extenuating circumstances can be imposed in a simulation study but require intense consideration

during the design construction phase to ensure they are appropriate in terms of real-world conditions.

Another major limitation to the study is that, while varying levels of standard error, bias, and RMSE resulted, these results do not show how decisions using the potentially flawed estimates would be affected. For example, the study shows that there may be a certain level of bias inherent in the estimate. However, the current study does not examine how this bias might affect decisions based off the estimates. If the method was being used to decide a plan of action regarding the individual (e.g., grade promotion, remedial level), the influence of the various approaches over this decision is unknown.

Other limitations of the study were the number of groups examined and the constraint of using fixed-length CATs, although these tests are often used in reality. Only two groups, a majority and minority group, were examined in this study. Results may vary when multiple groups exist in the population. For example, while the Population Composite prior was deemed most appropriate, this may not be the case when a large number of groups are used to obtain the composite mean. Also, the study used fixed-length termination rules. It would be interesting to see how a variable-length CAT functions in the design, specifically as it relates to how many items would need to be administered under each test prior and final ability estimate type to end the assessment, examining various termination rules.

Lastly, a limitation of the study is generalizability to other testing designs. The current study was designed to simulate testing in an education setting, such as high-stakes testing or formative/summative assessments. While it is possible that the study could be adapted for use in other domains, such as mental health, physical health, or personnel selection, studies would need to be conducted before application. Also, the design should not be discounted for use in other

domains because the recommendation of the current study is to stick with current administration techniques.

Recommendations for Future Research

The current study results, recommendations, and limitations do give rise to some possible avenues for continuing research.

- Since simulation studies allow for more control over the study design, another part of this current study could examine how the results would change when non-conforming individuals are introduced. These individuals would be those who do not perform as expected, even by the simulation algorithm. For example, while a 30-item test is not that long, a person might become fatigued towards the end of the exam and answer questions incorrectly that they otherwise would have successfully answered. The test priors may have a different influence on these individuals.
- While the current design may not recommend the use of group-based test priors for high-stakes testing, these test priors could be further examined in terms of formative assessments. These assessments can be used to hone in on a true group and thus may increase measurement precision.
- As discussed in the limitations, further research could examine the design in terms of variable-length CATs under various termination rules.
- Also, as discussed in the limitations, research could be conducted utilizing more than two groups. This may influence the results of the study, as well as recommendations made to test administrators.
- Currently, the study does not answer the question regarding how decisions based off the obtained estimates would be influenced. Therefore, an extension of the study should be

conducted in which classification accuracy should be examined. By placing a cutscore for a decision and investigating the relationship between classification based on the true and estimated ability, an understanding of how group-based priors affect decisions can be obtained.

- The current study examined only the Rasch and 2PL models. Two avenues of research could be aligned with this issue. One, a simulation could be conducted in which the Rasch model's item discrimination parameter is estimated but constrained to be the same across all items. Although direct comparisons would still be inappropriate, the 2PL model may still recover ability better since items have varying degrees of discriminatory values. However, the Rasch model is often used because it is easier to explain to the layperson. Also, the 3PL model might be utilized since the multiple-choice items often involve levels of guessing (i.e., lower asymptote).
- Multistage tests are becoming more common, since they offer many of the advantages of CATs plus others, such as item review within modules. The various test priors might be more beneficial in terms of these item-set adaptive tests. This avenue of research should be further investigated.

While the current results will not revolutionize the testing industry, they do give further validation to current test practices and do open more avenues of future research.

APPENDIX A
SIMULATION ONE CONVENTIONAL TEST ITEMS

Table A1. Item parameters for both IRT models for the 15-item test.

Item	Rasch Model		2PL Model	
	α	β	α	β
1	1	-1.843	1.769	-2.019
2	1	-1.413	1.711	-1.080
3	1	-0.938	2.136	-1.289
4	1	-1.315	1.529	-0.747
5	1	-0.898	1.924	-1.047
6	1	-0.361	1.816	-0.291
7	1	-0.195	2.134	-0.378
8	1	-0.028	3.287	0.316
9	1	0.372	1.641	-0.293
10	1	0.484	1.622	0.330
11	1	0.914	1.532	0.733
12	1	0.745	1.788	0.789
13	1	1.525	1.564	1.524
14	1	1.769	1.644	1.566
15	1	1.680	2.277	1.979

Table A2. Item parameters for both IRT models for the 30-item test.

Item	Rasch Model		2PL Model	
	α	β	α	β
1	1	-1.800	1.642	-1.852
2	1	-1.603	1.532	-1.619
3	1	-1.508	1.744	-1.437
4	1	-1.342	1.561	-1.523
5	1	-1.320	2.245	-1.270
6	1	-1.277	1.625	-0.977
7	1	-1.299	2.241	-0.614
8	1	-0.888	2.277	-0.899
9	1	-1.048	1.577	-1.059
10	1	-0.971	1.530	-0.877
11	1	-0.417	4.289	-0.534
12	1	-0.504	2.125	-0.723
13	1	-0.491	2.204	-0.629
14	1	-0.464	2.809	-0.283
15	1	-0.149	2.073	0.045
16	1	-0.137	1.843	-0.058
17	1	0.290	2.561	0.038
18	1	-0.100	1.647	0.140
19	1	0.641	2.669	0.603
20	1	0.277	1.992	0.264
21	1	0.590	2.591	0.982
22	1	0.845	2.144	0.597
23	1	1.094	2.870	1.227
24	1	0.648	1.540	1.139
25	1	1.387	1.509	1.413
26	1	1.135	1.543	1.314
27	1	1.465	2.048	1.455
28	1	1.647	3.081	1.609
29	1	1.622	1.923	1.536
30	1	1.722	1.767	1.655

APPENDIX B
SIMULATION TWO CONVENTIONAL TEST ITEMS

Table B1. Item parameters for both IRT models for the 15-item test.

Item	Rasch Model		2PL Model	
	α	β	α	β
1	1	-0.896	2.036	-0.935
2	1	-0.084	1.623	-0.560
3	1	-0.184	2.389	-0.463
4	1	-0.068	2.096	-0.365
5	1	0.158	2.176	-0.269
6	1	0.246	2.910	0.491
7	1	0.103	1.584	0.634
8	1	0.237	2.048	0.468
9	1	0.743	1.543	0.421
10	1	0.870	2.329	0.500
11	1	0.532	2.227	0.493
12	1	0.612	1.799	0.625
13	1	0.386	1.581	0.840
14	1	0.513	1.985	1.263
15	1	0.622	1.598	0.853

Table B2. Item parameters for both IRT models for the 30-item test.

Item	Rasch Model		2PL Model	
	α	β	α	β
1	1	-1.185	2.277	-0.899
2	1	-1.223	2.241	-0.614
3	1	-0.781	1.530	-0.877
4	1	-0.839	1.577	-1.059
5	1	-0.626	2.078	-0.659
6	1	-0.835	4.289	-0.534
7	1	-0.155	1.702	-0.730
8	1	-0.302	2.204	-0.629
9	1	-0.257	2.654	-0.229
10	1	-0.260	2.073	0.045
11	1	-0.194	2.637	0.303
12	1	0.071	2.561	0.038
13	1	0.308	1.561	0.254
14	1	0.495	2.669	0.603
15	1	0.434	2.227	0.493
16	1	0.397	2.606	0.533
17	1	0.509	1.576	0.512
18	1	0.651	1.916	0.598
19	1	1.082	2.445	0.396
20	1	0.748	1.584	0.634
21	1	0.432	2.048	0.468
22	1	0.534	1.887	0.723
23	1	0.601	3.771	0.571
24	1	0.629	2.144	0.597
25	1	1.011	2.591	0.982
26	1	1.207	4.398	0.968
27	1	1.300	2.870	1.227
28	1	1.244	1.607	1.245
29	1	1.672	1.509	1.413
30	1	1.330	1.775	1.277

REFERENCES

Babcock, B., & Weiss, D. J. (2012). Termination criteria in computerized adaptive tests: Do variable-length CATs provide efficient and effective measurement? *Journal of Computerized Adaptive Testing, 1*(1), 1-18.

Bergstrom, B. A., Lunz, M. E., & Gershon, R. C. (1992). Altering the level of difficulty in computer adaptive testing. *Applied Measurement in Education, 5*(2), 137-149.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Blackburn, M. L. (2004). The role of test scores in explaining race and gender differences in wages. *Economics of Education Review, 23*, 555-576.

Bock, R. D., Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443-459.

Breithaupt, K., Ariel, A., & Veldkamp, B. P. (2005). Automated simultaneous assembly for multistage testing. *International Journal of Testing, 5*(3), 319-330.

Carroll, J. B. (1993). *Human Cognitive Abilities: A Survey of Factor-Analytic Studies.* Cambridge, England: Cambridge University Press.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54*(1), 1-22.

Chang, H. H. (2014). Psychometrics behind computerized adaptive testing. *Psychometrika*, DOI: 10.1007/SI1336-014-9401-5.

Chuah, S. C., Drasgow, F., & Luecht, R. (2006). How big is big enough? Sample size requirements for CAST item parameter estimation. *Applied Measurement in Education, 19*(3), 241-255.

Crocker, L., & Algina, J. (2008). *Introduction to Classical & Modern Test Theory.* Mason, OH: Cengage Learning.

Cronbach, L. J., & Gleser, G. C. (1965). Two-stage sequential selection. In *Psychological Tests and Personnel Decisions*. Urbana, IL: University of Illinois Press.

de Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory.* New York, NY: The Guilford Press.

Doebler, A. (2012). The problem of bias in person parameter estimation in adaptive testing. *Applied Psychological Measurement, 36*(4), 255-270.

du Toit, M. (Ed.). (2003). *IRT from SSI. BILOG-MG, MULTILOG, PARSCALE, TESTFACT.* Lincolnwood, IL: Scientific Software International.

Eggen, T. J. H. M. (2011). Computerized classification testing with the Rasch model. *Educational Research and Evaluation, 17*(5), 361-371.

Eitelberg, M. J. (1981). *Subpopulation differences in performance on tests of mental ability: Historical review and annotated bibliography*. (HumRRO-TM-81-3). Washington, DC: Office of the Assistant Secretary of Defense (Manpower Reserve Affairs and Logistics).

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

ETS. (2016). *A snapshot of the individuals who took the GRE revised general test.* Princeton, NJ: Author. Retrieved from https://www.ets.org/s/gre/pdf/snapshot_test_taker_data_2015.pdf

Garrison, W. M., & Baumgarten, B. S. (1986). An application of computer adaptive testing with communication handicapped examinees. *Educational and Psychological Measurement, 46*(1), 23-35.

General Accounting Office. (1990). *Military training: Its effectiveness for technical specialities is unknown.* (GAO Code 973276) OSD Case 8371.

Gershon, R. C. (2005). Computer adaptive testing. *Journal of Applied Measurement, 6*(10, 109-127.

Gnambs, T., & Batinic, B. (2011). Polytomous adaptive classification testing: Effects of item pool size, test termination criterion, and number of cutscores. *Educational and Psychological Measurement, 71*(6), 1006-1022.

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.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*(4), 347-360.

Guilford, J. P. (1954). *Psychometric Methods.* New York, NY: McGraw-Hill Book Company, Inc.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38-47.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publications, Inc.

Hambleton, R. K., van der Linden, W.J., & Wells, C. S. (2010). IRT models for the analysis of polytomously scored data: Brief and selected history of model building advances. In M. L. Nering, R. Ostini (Eds.), *Handbook of Polytomous Item Response Theory Models* (21-42). New York, NY: Routledge/Taylor & Francis Group.

Hambleton, R. K., & Xing, D. (2006). Optimal and nonoptimal computer-based test designs for making pass-fail decisions. *Applied Measurement in Education, 19*(3), 221-239.

Han, K. T. (2013). MSTGen: Simulated data generator for multistage testing. *Applied Psychological Measurement, 37*(8), 666-668.

Harris, D. (1989). Comparison of 1-, 2-, and 3-parameter IRT models. *Educational Measurement: Issues and Practice, 8(*1), 35-41.

Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice, 26*(2), 44-52.

Horn, J. L. & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology, 57*(5), 253-270.

Huynh, H., & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics, 1*(1), 69-82.

Lathrop, Q. N., & Cheng, Y. (2013). Two approaches to estimation of classification accuracy rate under item response theory. *Applied Psychological Measurement, 37*(3), 226-241.

Lee, W-C. (2010). Classification consistency and accuracy for complex assessments using item response theory. *Journal of Educational Measurement, 47*(1), 1-17.

Lord, F. M. (1971). A theoretical study of two-stage testing. *Psychometrika, 36*(3), 227-242.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Hillsdal, New Jersey: Lawrence Erlbaum Associates.

Lord, F. M. (1983). Unbiased estimators of ability parameters of their variance, and of their parallel-forms reliability. *Psychometrika, 48*(2), 233-245.

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley Publishing Company.

Luecht, R. M. (2006). Operational Issues in Computer-Based Testing. In D. Bartram, R. K. Hambleton (Eds.) , *Computer-based testing and the Internet: Issues and advances* (pp. 91-114). New York, NY US: John Wiley & Sons Ltd.

Luecht, R. (2014). Design and implementation of large-scale multistage testing systems. In D. Yan, A. A. von Davier, C. Lewis (Eds). *Computerized Multistage Testing: Theory and Applications* (69-84). Chapman and Hall/CRC.

Luecht, R., Brumfield, T., & Breithaupt, K. (2006). A testlet assembly design for adaptive multistage tests. Applied Measurement in Education, 19(3), 189-202.

Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement, 35*(3), 229-249.

Lynn, R. (2006). *Race Differences in Intelligence: An Evolutionary Analysis.* Augusta, GA: Washington Summit Publishers.

Lynn, R. & Kanazawa, S. (2011). A longitudinal study of sex differences in intelligence at ages 7, 11 and 16 years. *Personality and Individual Differences, 51*, 321-324.

Magis, D., & Raíche, G. (2012). Random Generation of Response Patterns under Computerized Adaptive Testing with the R Package catR. *Journal of Statistical Software, 48*(8), 1-31. URL: http://www.jstatsoft.org/v48/i08/.

Matteucci, M., Mignani, S., & Veldkamp, B. P. (2011). Prior distributions for item parameters in IRT models. *Communications in Statistics – Theory and Methods, 41*, 2944-2958.

Matteucci, M. & Veldkamp, B. P. (2011). Including empirical prior information in test administration. In B. Fichet, D. Piccolo, R. Verde, & M. Vichi (Eds.), *Classification and Multivariate Analysis for Complex Data Structures* (173-181). Berlin: Springer.

Matteucci, M. & Veldkamp, B. P. (2013). On the use of MCMC computerized adaptive testing with empirical prior information to improve efficiency. *Statistical Methods & Applications, 22*(2), 243-267.

Matteucci, M. & Veldkamp, B. P. (2015). The approach of power priors for ability estimation in IRT models. *Quality & Quantity, 49*(3), 917-926.

Mead, A. D. (2006). An introduction to multistage testing. *Applied Measurement in Education, 19*(3), 185-187.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*(3), 449-458.

Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement, 23*(3), 187-194.

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

Mislevy, R. J. & Sheehan, K. M. (1989). The role of collateral information about examinees in item parameter estimation. *Psychometrika, 54*(4), 661-679.

Nydick, S. W. (2014). The sequential probability ratio test and binary item response models. *Journal of Educational and Behavioral Statistics, 39*(3), 203-230.

Patsula, L. N. (1999). *A comparison of computerized adaptive testing and multi-stage testing.* Retrieved from ProQuest Digital Dissertations (AAT 9950199).

Patton, J. M., Cheng, Y. Yuan, K., & Diao, Q. (2013). The influence of item calibration error on variable-length computerized adaptive testing. *Applied Psychological Measurement, 37*(1), 24-40.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: http//www.R-project.org/.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21*(1), 25-36.

Reckase, M. D. (2009). *Multidimensional Item Response Theory* (1-232). New York, NY: Spriner.

Roberts, J. S., & Thompson, V. M. (2011). Marginal Maximum A Posteriori item parameter estimation for the Generalized Graded Unfolding Model. *Applied Psychological Measurement, 35*(4), 259-279.

Robin, F, Steffen, M., & Liang, L. (2014). The multistage test implementation of the GRE Revised General Test. In D. Yan, A. A. von Davier, C. Lewis (Eds). *Computerized Multistage Testing: Theory and Applications* (325-342). Chapman and Hall/CRC.

Roth, P. L., Bevier, C. A., Bobko, P., Switzer III, F. S., & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. Personnel Psychology, 54, 297-330.

SAT. (2015*). 2015 college-bound seniors: Total group profile report.* CollegeBoard. Retrieved from https://secure-media.collegeboard.org/digitalServices/pdf/sat/total-group-2015.pdf

Smith, R., & Lewis, C. (2014). Multistage testing for categorical decisions. In D. Yan, A. A. von Davier, C. Lewis (Eds). *Computerized Multistage Testing: Theory and Applications* (189-204). Chapman and Hall/CRC.

Spearman, C. (1904). "General Intelligence," objectively determined and measured. The American Journal of Psychology, 15(2), 201-292.

Spearman, C. (1927). *The Abilities of Man: Their Nature and Measurement.* London: Macmillan.

Stone, C. A., Weissman, A., & Lane, S. (2005). The consistency of student proficiency classifications under competing IRT models. *Educational Assessment, 10*(2), 125-146.

Thissen, D., & Mislevy, R. J. (2010). Testing algorithms. In H. Wainer (Ed.), *Computerized Adaptive Testing: A Primer, Second Edition* (101-134). New York, NY: Routledge.

Thompson, N. A. (2007). A practitioner's guide for variable-length computerized classification testing. *Practical Assessment, Research, & Evaluation, 12*(1), 1-13.

Thompson, N. A. (2009). Item selection in computerized classification testing. *Educational and Psychological Measurement, 69*(5), 778-793.

Thompson, N. A. (2011). Termination criteria for computerized classification testing. *Practical Assessment, Research, & Evaluation, 16*(4), 1-7.

Thompson, N. A., & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation, 16*(1). Available online: http://pareonline.net/getvn.asp?v=16&n=1.

U.S. Census Bureau. (2016). Annual estiamtes of the resident population by sex, race, and Hispanic origin for the United States, States, and counties: April 1, 2010 to July 1, 2015. U.S. Census Bureau, Population Division.

van der Linden, W. J. (1999). Empirical initialization of the trait estimator in adaptive testing. *Applied Psychological Measurement, 23*(1), 21-29.

van der Linden, W. J., Entink, R. H. K., & Fox, J-P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement, 34*(5), 327-347.

van der Linden, W. J., & Pashley, P. J. (2010). Item selection and ability estimation in adaptive testing. In W.J. van der Linden, C.A.W. Glas (Eds.). *Elements of Adaptive Testing* (3–30). Springer.

Veerkamp, W. J. J. & Berger, M. P. F. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics, 22*(2), 203-226.

Veldkamp, B. P. & Matteucci, M. (2013). Bayesian computerized adaptive testing. *Ensaio: Avalição e Políticas Públicas em Educação, 21*(78), 57-81.

Wainer, H., & Eignor, D. (2010). Caveats, pitfalls, and unexpected consequences of implementing large-scale computerized testing. In H. Wainer (Ed.) *Computerized Adaptive Testing: A Primer, Second Edition* (271-299). New York, NY: Routledge.

Wang, T., Hanson, B. A., & Lau, C. A. (1999). Reducing bias in CAT trait estimation: A comparison of approaches. *Applied Psychological Measurement, 23*(3), 263-278.

Wang, S., & Wang, T. (2001). Precision of Warm's Weighted Likelihood Estimates for a polytomous model in computerized adaptive testing. *Applied Psychological Measurement, 25*(4), 317-331.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427-450.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*(4), 473-492.

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology, 53*(6), 774-789.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*(4), 361-375.

Weiss, D. J., & Vale, C. D. (1987). Adaptive testing. *Applied Psychology, 36*(3-4), 249-262.

Weissman, A. (2014). IRT-based multistage testing. In D. Yan, A. A. von Davier, C. Lewis (Eds). *Computerized Multistage Testing: Theory and Applications* (152-168). Chapman and Hall/CRC.

Wise, L., Welsh, J., Grafton, F., Foley, P., Earles, J., Sawin, L., & Divgi, D. R. (1992). *Sensitivity and fairness of the Armed Services Vocational Aptitude Batterty (ASVAB) technical composites.* Seaside, CA: Defense Manpower Data Center.

Xing, D. & Hambleton, R. K. (2004). Impact of test design, item quality, and item bank size on the psychometric properties of computer-based credentialing examinations. *Educational and Psychological Measurement, 64*(1), 5-21.

Yan, D., Lewis, C., & von Davier, A. A. (2014). Overview of computerized multistage tests. In D. Yan, A. A. von Davier, C. Lewis (Eds). *Computerized Multistage Testing: Theory and Applications* (3-20). Chapman and Hall/CRC.

Yi, Q., Wang, T., & Ban, J. C. (2001). Effects of scale transformation and test-termination rule on the precision of ability estimation in computerized adaptive testing. *Journal of Educational Measurements, 38*(3), 267-292.

Zenisky, A. L, & Hambleton, R. K. (2014). Multistage test designs: moving research results into practice. In D. Yan, A. A. von Davier, C. Lewis (Eds). *Computerized Multistage Testing: Theory and Applications* (21–38). Chapman and Hall/CRC.

Zimowski, Muraki, Mislevy, & Bock. (2003). *BILOG-MG 3: Item analysis and test scoring with binary logistic models.* Chicago, IL: Scientific Software. [Computer software].

Zwick, R., & Bridgeman, B. (2014). Evaluating validity, fairness, and differential item functioning in multistage testing. In D. Yan, A. A. von Davier, C. Lewis (Eds). *Computerized Multistage Testing: Theory and Applications* (271-284). Chapman and Hall/CRC.