COMPARABILITY OF EXAMINEE PROFICIENCY SCORES ON COMPUTER ADAPTIVE TESTS USING REAL AND SIMULATED DATA

by

JOSIAH JEREMIAH EVANS

A Dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Educational Psychology

Written under the direction of

Jimmy de la Torre, Ph.D.

And approved by

____

____

____

____

New Brunswick, New Jersey

January, 2010

ABSTRACT OF THE DISSERTATION

Comparability of Examinee Proficiency Scores on Computer-Adaptive Tests Using Real and Simulated Data

By JOSIAH JEREMIAH EVANS

Dissertation Director:
Jimmy de la Torre, Ph.D.

In measurement research, data simulations are a commonly used analytical technique. While simulation designs have many benefits, it is unclear if these artificially generated datasets are able to accurately capture real examinee item response behaviors. This potential lack of comparability may have important implications for administration of computer adaptive tests (CAT) which display proficiency-targeted items to examinees. In addressing this problem, this study sought to compare results from real testing data to that of simulated data to determine the extent to which simulated data are an accurate representation of real-world testing data. Specifically, this study matched real examination data from multiple administrations of the Law School Admission Test to create a single large dataset with 534 items and 5,000 synthetic examinees. From this dataset examinee proficiency estimates and item parameters were obtained, which were used to create 100 simulated item response datasets. Both real and simulated data were utilized in two post-hoc testing formats: CAT and linear format examinations. The CAT administrations used the item-level adaptive method; the linear tests were constructed by selecting items using stratified random sampling. In addition to the two data types and two test administration formats, the impact of three varying test lengths (25, 35, and 50 items) on proficiency estimation was examined. For linear tests, results demonstrated that replication of original proficiency estimates from simulated data was variable, depending on test length, items selected, and examinee proficiency levels. Randomly constructed linear tests with extreme item parameter values resulted in test instability which yielded less accurate proficiency recovery. For most datasets, CAT format tests yielded improved true proficiency recovery as compared to their linear test counterparts. Generally, the longest length 50-item CAT simulated data tests yielded the best replication of original real data proficiency estimates. CAT format tests performed well given real or simulated data, whereas linear tests displayed more performance variation compared to their CAT counterparts. The tails of the distributions showed the greatest variation between data types and conditions. The results of this dissertation support the use of simulated data when the items used to construct the tests reflect non-extreme item parameter values.

Acknowledgements

Though there are many people deserving thanks for facilitating my completion of this dissertation, I shall attempt to be brief here in my acknowledgements of these contributions. Firstly, I would like to thank my advisor, Jimmy de la Torre for his considerable help and assistance in many forms. I would also like to thank my remaining committee members, Douglas Penfield, Gregory Camilli and Peter Pashley for their help and advice. I would especially like to thank Peter Pashley for the idea for this dissertation as well as for his facilitating the resources necessary to get the appropriate LSAT data. Besides those formally on my committee, I would also like to thank a few others who helped me to move along this dissertation process. My supervisor at LSAC, Lynda Reese deserves many thanks for her patience and flexibility throughout this process, as well as for her moral support. I would also like to thank Susan Dalessandro at LSAC for her assistance with obtaining the correct datasets. When it comes to moral support, I have been very fortunate to have many supportive friends, colleagues, and family members, too many to name here. In terms of friends and colleagues, I would like to extend my thanks to a select few not yet named: Merav (Pfeffer) Dechaume, Adam Prowker, Anne Gallagher, Andrea Thornton Sweeney, Debbie Suto, Arlene Amodeo, Michele Lawrence, Alex Weissman, Dmitry Belov, and Weiling Deng. More generally, I would like to thank my employers, the American Institute of Certified Public Accountants, and the Law School Admission Council for their generous financial support of my education. Broadly, I would also like to thank the faculty, staff, and administration of Rutgers University for their hard work, assistance, and diligence. Of particular note, on the Graduate School of Education staff, I would like to thank Kris Spaventa for always going

the extra mile to help me manage all the administrative issues inherent in completing a Ph.D., particularly while I was working many miles from the Rutgers campus.

No dissertation would be possible without the support of family members, who deserve a special place for the sacrifices they were willing to make to ensure I could get to this point. Firstly, I would like to thank my partner and confidant, Richard Barry, for all the sacrifices and assistance he gave me over the past thirteen years of my graduate education. There is little that, at some point, does not suffer in a family when one member is working while attending university and I am grateful for the understanding extended to me by my family. I would also like to thank my daughter, Abigail, for her understanding when I could not attend all of her events due to classes, exams, or other academic concerns. Few people are able to get to the point of earning a doctorate without having a supportive family of origin, and my experience is no exception. I have been fortunate indeed to have wonderful parents, grandparents, and siblings who were all supportive in their own myriad ways. Though they may not have always understood this academic journey, they always extended their support, and for that I am grateful indeed. In addition to my own birth family, I would like to thank my partner's family, The Barrys, for their support and encouragement as well as their accepting me into their family in a warm and considerate manner. Though I am hard pressed to single out any of my family members as being more instrumental than another, I would like to extend special thanks to my mother and her parents, my grandparents, for their financial support and thoughtfulness over the years, despite the sacrifices they faced to do so. As a teacher, my mother always knew the value of education and encouraged me to pursue it as far as possible, despite the sacrifices. Well, I made it, mom – I went as far as I could and made it to the top! I sincerely doubt I would have made it this far had it not been for my mother's insistence on the importance of education. Thank you, mom. On a related note, I would like to thank the teachers of America for their dedication to the art and science of teaching.

Dedication

I would like to dedicate this dissertation to my beloved grandfather, Raymond Cornwell Thomas, who departed from this mortal coil while I was writing this dissertation. Pop, you will always be remembered very fondly by your loving family, friends, and any who were touched by your kind and gentle spirit.

Table of Contents

List of Tables

List of Figures

CHAPTER I. INTRODUCTION

Educational measurement is concerned with the assessment of knowledge or skills of individuals. Typically, many such examinees experience this assessment in the form of a large-scale, standardized tests such as the Scholastic Aptitude Test, or SAT®. Large-scale standardized assessments are utilized for many purposes, from college entrance to skills assessment to licensure examinations. With the proliferation of many professions requiring specialized skills, the popularity of standardized examinations has been increasing and shows few signs of slowing. As demand for educational measurement products grows, the need for sophisticated, targeted assessments, such as computer-adaptive testing (CAT) methods grows. Moreover, data simulations are often used as tools to meet the need for timely, ethical research on measurement models and testing methods. What is unclear, however, is how representative these simulation studies are of real examinee responses, particularly as research programs transition from traditional paper-and-pencil (P&P) tests to advanced CAT designs.

1.1 Computer-Adaptive Testing

Many recent educational measurement studies focus on adaptive testing issues and numerous books deal with the adaptive testing concerns (e.g., Parshall, Spray, Kalohn, & Davey, 2002; Wainer, 2000). A recent search of the ERIC database for “adaptive testing” found 900 references spanning 40 years (retrieved September 12, 2009 from Rutgers University Libraries ERIC database). Clearly, adaptive testing has been heavily researched and is an important and timely topic for measurement professionals.

Reflecting the compelling impact of these numerous research studies, a number of prominent testing programs have converted to, or are in the process of converting to, computer based tests with adaptive features of some kind (e.g., the Uniform CPA Examination, The General Management Admission Test, Test of English as a Foreign Language). Item response theory (IRT) is designed to relate an examinee's proficiency to a given item by way of a probabilistic function. As Hambleton, Swaminathan, and Rogers (1991) explained, item response models function using both examinee proficiency (theta, symbolically θ) and the item parameters. Item response theory provides a psychometric framework for CAT. The link between IRT and adaptive testing is quite strong. The relationship is so strong that Wainer (2000) stated that "many believe that adaptive testing is the *raison d' etre* of IRT" (p. 9). There are numerous ways to achieve the adaptive function, from item-level adaptation to multi-stage adaptive testing (Thissen & Mislevy, 2000). Computer-adaptive testing (CAT) utilizes item response theory (IRT) to determine many aspects of the examinee assessment process from item selection to proficiency estimation. Typically, items with known parameters are administered in an adaptive fashion; items are selected based on an examinee's demonstrated proficiency estimate derived from previously administered items.

1.2 Data Simulations

In quantitative research studies, data simulations are an invaluable tool with origins in the natural sciences (Metropolis & Ulam, 1949). Simulations can serve to inform theory, research, practice, and other areas. In educational measurement research, simulation studies are ubiquitous and influential (Harwell, Stone, Hsu, & Kirisci, 1996). There is little published research, however, demonstrating the validity of using these simulations to represent real examinee item response behavior, as collected in real-world testing situations.

In educational measurement research, data simulations have become commonplace. A recent search for the term “simulations” on the ERIC database yielded over 3,700 hits (Rutgers University ERIC database, September 12, 2009). In addition, during the period of 1994-1995, Harwell, Stone, Hsu, and Kirisci (1996) found that in a content analysis of three prominent measurement journals (*Applied Psychological Measurement*, *Psychometrika*, and *Journal of Educational Measurement*), nearly one-third of all articles utilized simulation methods in their research. As noted previously, the prevalence of simulation studies has increased. Recent examples include differential item functioning (DIF) studies (Schnipke, Roussos, & Pashley, 2000; Zwick & Thayer, 2003), test security (McLeod, Lewis, & Thissen, 1999; Wen, Chang, & Hau, 2000), parameter recovery (Evans & Weissman, 2005), scaling and calibration (Ban, Hanson, Yi, & Harris, 2002), item answer alteration policies (Bowles & Pommerich, 2001), item exposure control methods (Chang & Twu, 2001), cognitive diagnosis (de la Torre, 2009; de la Torre & Douglas, 2008) and many more.

Stated briefly, a data simulation study is a statistical sampling procedure in which the researcher generates output stochastically. Typically, these studies are used to address a research question; therefore, a model exists for generating the numbers and subsequently using the results to test hypotheses. One type of popular simulation technique is the Monte Carlo (MC) method. This paper will not address MC methods specifically, as they will be considered as a special subset of the more generalized simulation methods.

Several authors have posited that simulation studies should be held to the same standard as real-world empirical studies, deserving the same attention to detail and experimental design (Harwell, Stone, Hsu, and Kirisci, 1996; Spence, 1983). Harwell, Stone, Hsu, and Kirisci discussed published Monte Carlo simulation IRT studies and the potentially significant problems that can arise from imprecise simulation design. They noted that “[o]ne limitation of these [Monte Carlo simulation] studies is that the usefulness of the results is highly dependent on how realistic the conditions modeled are.” (p. 104). Studies with unrealistically modeled conditions are unlikely to fully represent real-world responses of examinees.

1.3 Simulated Data Using CAT Designs

Simulation methods are commonplace and their results are typically taken as valid evidence for making decisions about testing programs, with real consequences for examinees and other stakeholders. Few published research articles to date have utilized real examinee data in creating simulations for adaptive test comparisons. Therefore, this study seeks to utilize real test data in a CAT design, which will reflect the complexity of real-world CAT estimation. A main focus of many test designs is to obtain accurate proficiency estimates. Studying this comparison between simulated and real data in adaptive test designs will allow the researcher to understand the level of generalizability of their simulation studies to real-world testing conditions. Without this knowledge, testing programs will continue to draw potentially erroneous conclusions about the results of simulation studies designed to study computer-adaptive examination outcomes.

Purpose

This study proposes to examine the comparability of simulation data as compared to real examinee responses from a large-scale, paper-and-pencil (P&P) examination used in a computer-adaptive testing format. The purpose of this study is not to evaluate potential improvements to IRT model fit; rather, the purpose is to describe the extent to which CAT simulation results capture real-world item responses.

The Law School Admission Test (LSAT) Logical Reasoning section data were used to create a large dataset of real synthesized examinee responses with 534 items and 5,000 examinees. These data were exclusively multiple-choice, dichotomously scored items. Using the proficiency and item parameters obtained from the real data, simulated datasets were created. Both real and simulated data were utilized in creating linear, paper-and-pencil (P&P) type tests as well as item-level CATs. The primary outcome variable for this study was $\hat{\theta}$ (estimated proficiency). In addition, for P&P tests, classical test theory indices of item difficulty, and item discrimination were obtained. For CATs, item exposure rates were summarized. Examinee characteristics, specifically the proficiency estimates, were compared across conditions using Pearson correlation, bias, root mean squared error, standard error, and relative efficiency. Results from this study will provide a realistic analysis of the effects of using simulations to mimic real examinee item response behaviors.

CHAPTER II. RELEVANT LITERATURE

First, it will be instructive to review the basics of IRT and simulation techniques, and how they work together to simulate examinee item responses. Some published IRT simulation studies have been poorly designed and therefore may not validly support the authors' conclusions (Davey, Nering, & Thompson,1997). Thus, some research studies have found that simulated IRT data may fail to properly replicate the complexity of real examinee data. This section explores both the inadequacy of many simulation designs, as well as the inadequate modeling inherent in simulation research. Specifically, this chapter will discuss the basic theory behind IRT-based simulations, the limitations of published studies, and the limitations of the models for accurately recovering all response characteristics within real test data.

2.1 Item Response Theory

Before discussing the issues leading to the reasoning behind this study, it may be useful to review the basic concepts of the traditional IRT three-parameter logistic model (3PL). Below is the typical 3PL model, as described in Hambleton, Swaminathan, and Rogers (1991). It includes discrimination (*a*), difficulty (*b*), and pseudo-guessing (*c*) parameters as well as proficiency (θ) and is represented mathematically as

$$P_i(\theta) = c_i + (1 - c_i)\ [e^{Da_i(\theta - b_i)} / 1 + e^{Da_i(\theta - b_i)}]\ (i\text{=1, 2, … n}).$$

where $P_i(\theta)$ is the probability that a randomly chosen examinee with proficiency θ will answer item *i* correctly; *e* is Euler's number (2.718); *D* is 1.7. Detailed information can be found in Hambleton et al. (1991), Lord (1980), and others.

2.1.1 IRT Simulations

Simulation studies are ubiquitous in many research fields due to the utility and simplicity of these designs. In IRT-based educational measurement research, many studies rely on simulation methods for varied purposes, such as demonstrating a particular statistical model's robustness to violations of assumptions, modeling examinee behavior on a given test, studying effects of applied psychometric methods and others. If, however, these simulations are unable to create data that accurately represent real-world examinee response patterns, the outcomes of these simulation studies are questionable.

Given the proliferation of simulation designs, it is reasonable to give only a brief overview of the generalized method. Data simulation methods permit researchers to empirically examine a given characteristic using random samples drawn from known populations (e.g., Mooney, 1997). In educational measurement theory, these simulation studies are commonplace and are typically assumed to be valid representations of the phenomenon of interest (Davey, Nering, & Thompson, 1997). As noted previously, a large number of studies utilizing IRT-based data simulation methods have been published or presented in recent decades, covering many topics from item parameter estimation and recovery (e.g., Harwell & Janosky, 1991; Hulin, Lissak, & Drasgow, 1982; McCauley & Mendoza, 1985), to dimensionality (e.g., Ansley & Forsyth, 1985; De Ayala, 1994), test equating (e.g., Prowker, 2005; Fairbank, 1985), and proficiency estimation (e.g., Kim & Plake, 1993; Yen, 1987), as well as many others.

IRT-based simulation designs typically include a basic set of stochastic methods facilitating researcher's goals, as summarized in Davey, Nering, and Thompson (1997):

1. Specify the form of the IRT model. Typically, these include such components as dimensionality, independence, cumulative distribution functions, and others.
2. Specify parameters of IRT model, either by simulating or selecting operational items from a real examination.
3. Specify the form of the examinee proficiency (i.e.,θ) distribution

The actual item responses are simulated by randomly drawing a θ value from the designated distribution which is used in the computation of a probability value using the IRT model (such as 3PLM) and the item parameters at the selectedθ. This computed value is compared to a random value from (0, 1); if the random uniform value is less than the computed probability value, it is scored as correct; if the random uniform value is greater than the computed probability, it is scored as incorrect. The above method is the basic form to which this paper refers.

2.1.2 Limitations of IRT Simulations

Using established techniques, simulations are often presumed to be valid representations of real-world test data (Davey, Nering, & Thompson, 1997). Employing traditional data distributions such as the uniform and normal distributions, as well as elements of other theoretically accepted measurement models, such as IRT, many researchers and practitioners assume that simulation studies create data which lead to valid conclusions. Little empirical evidence exists, however, to suggest that such simulation results accurately represent real-world examinee testing behavior.

An analysis by Harwell et al. (1996) summarized many IRT simulation studies and analyzed the research methods and conclusions. Harwell et al. state that there are few resources that focus on the rigor of these simulation studies or how to properly conduct these studies. In fact, by the 1970s, some amount of incredulity had arisen from publications presenting simulation studies with less than ideal designs. To reduce the number of poorly constructed studies, Harwell et al. state that there were two major publications created to curtail the number of inadequate simulation studies: An article by Hoaglin and Andrews (1975), and an article by the Psychometric Society in *Psychometrika* (1979) provided clear explanation of acceptable criteria for MC simulation designs. Both articles gave similar standards for appropriate utilization and design of these studies. As noted by Harwell et al. in 1996, the trend of publishing simulation studies using inadequate designs continued many years after the publication of these articles. They state that many IRT simulation studies are conducted improperly or are of insufficient quality to warrant valid conclusions. For example, in 26 parameter estimation studies they evaluated, they found that most studies failed two or more of the standards set forth by the Psychometric Society or by Hoaglin and Andrews. They concluded that, while parameter estimation studies performed poorly, dimensionality studies performed even more poorly.

As noted previously, simulations are based on accepted psychometric models using standard techniques which give the appearance of validity. The ideal-world model of these simulation designs, however, does not necessarily accurately capture true real-world examinee behaviors. Examinee response behavior is a complex process, yet educational measurement models typically fail to incorporate this complexity, as noted in Davey, Nering, and Thompson (1997). They have shown that less than 50% of the score variance is accounted for by $\hat{\theta}$. Snow and Lohman (1989) have noted that the psychological processes underlying item responding are varied and complex and, therefore, are far from simplistic processes; unfortunately, theoretical models may imply such simplicity inappropriately. Snow and Lohman observe that measurement models do not adequately capture this complexity. As such, the researcher can infer that inadequate measurement models create inadequate simulations as a result of using these models.

Given that only a small proportion of the variance in test scores is accounted for by $\hat{\theta}$, it is reasonable to assume that there may be additional factors influencing examinee test response behaviors. In turn, these additional factors influence item response modeling, a key component used in simulations. Nevertheless, many simulation studies in psychometrics have relied simply on the IRT model, and on potentially unrealistic data distributions. Therefore, some simulation studies produce valid results and others do not. While this ambiguity could prove confusing to the reader, some resolution may be obtained by the level-of-analysis concept: In some studies, the measures used and statistics employed may reveal differences between models, while other methods and analyses fail to reveal important differences, despite using the same models.

2.1.3 Addressing IRT Simulation Limitations

One method of examining simulation models is to create augmented models to improve the fit of simulated data to real data. Augmenting measurement models with additional information may yield improved recovery of real examinee item responses. Davey, Nering, and Thompson (1997) found that by using multidimensional IRT (MIRT), they were able to more accurately replicate real examinee response outcomes than they could by using simple unidimensional IRT. Davey et al. objected to violations of assumptions used in many published MC studies, citing three major, but often unsupported, assumptions. First, item score regressions on proficiency are assumed to be logistic, with monotonically increasing functions. Unfortunately, this assumption may be incorrect, as nonmonotonically increasing functions have been found in research on this topic (Levine, 1984). Davey et al. observed that a second problem with simulations is that item responses are assumed to be determined only by latent proficiency (θ), but this assumption has been determined to be inaccurate. As noted earlier, even on well-constructed tests, proficiency typically accounts for less than half of score variance. The third issue the authors address is that simulated item parameters used in response generation are unlikely to resemble real item parameters taken from actual examinations, and the misspecification can be substantial.

Despite these issues, Davey, Nering, and Thompson conclude that unidimensional data simulation is sufficiently similar to real data in several crucial ways: item passing rates, item test score correlations, number right score distributions, and test reliabilities. They state that typical unidimensional simulations may correspond to real data on the aforementioned traits, but they also say that more specialized analyses of interactions and other features may reveal nontrivial differences between the two data sources. Unfortunately, it seems the authors neither indicated any literature references nor clearly indicated specific empirical analyses used to support these claims. Thus, the veracity of these claims is uncertain. Nonetheless, Davey et al. found some potential support for simulation’s comparability to real data for select features. Their complex multidimensional modeling, however, was able to capture the full battery of real examinee item response characteristics, leading the reader to conclude that simple unidimensional models have mixed capacity to mimic real-world item response behaviors.

Using a related idea to that of Davey et al., Stocking, Steffen, and Eignor (2001) found that augmented simulation models when compared to real data revealed differences at the detailed analysis level, but failed to reveal differences at the aggregate level**.** Similar to Davey et al.'s approach, Stocking et al. also used a modeling augmentation technique to improve simulations to more accurately reflect real examinee behavior. Unlike Davey et al., however, Stocking et al. attempted to model real data based on an a priori construct of modeling missing responses. Stocking et al. proposed that these missing responses were the result of two situations: items not being reached and examinees guessing at random. They call their model the Test Taker Model (TTM), and it includes modeling of guessing at random responses and nonresponses of items not reached by the examinee.

The authors discovered that discrete items yielded less guessing at random than did the set-based items. They also found that those with the lowest proficiency estimates were more likely to finish the test versus those with high proficiency estimates. Unfortunately, they also found that their TTM based analyses had lower reliability than the more parsimonious 3PL modeling. The authors propose that the lower reliability may have been the result of within-examinee behavior patterns such as dependencies between items on guessing at random behavior. These dependencies were not modeled, and the authors suggested that perhaps guessing behaviors may be an intra-examinee effect, similar to a personality trait. The authors found that their TTM worked well at modeling items not reached, with more accurate modeling than is accomplished using traditional multidimensional IRT (MIRT) methods. Finally, the authors noted that the similarity between simulations modeled with traditional 3PL and their more complex TTM model was high. They stated that the 3PL is sufficient for modeling simulation data and that more complicated methods may not yield more accurately simulated data in every case.

Proposing a theoretical basis for augmented modeling, Stocking, Steffen, and Eignor’s TTM model improved upon real data model fitting similarly to the less theoretical modeling shown in Davey et al. (1997). While Davey et al. chose to use high dimensionality modeling with no a priori conceptualization, Stocking et al. chose to explicitly model two additional constructs which they assert influenced test taker response patterns. While they had some success particularly with not-reached item responses, they did not find substantial improvements using a more complex simulation model over the more traditional 3PL model. Thus, the IRT modeled simulations may accurately reproduce statistics for some basic analyses, but differences may be obtained when analyses are targeted at a finer level of detail.

2.2 Modeling Cognitive Complexity

How these augmented IRT models were able to capture the additional information contained in real data is unclear. The Davey et al. paper did not attempt to delineate what their enhanced model captured. They only sought to demonstrate that a MIRT type model could better replicate real data. Stocking et al. did seek to specify two critical respondent behaviors which they believed impacted the real data simulation. While they were successful at the finer level of analysis, they failed to note any differences at the grosser level, such as overall scaled scores. Therefore, the researcher is left to wonder what unmodeled constructs are hidden in real data that are not being captured by simulated data. Snow and Lohman (1993) state that educational psychometric measurement (EPM) fails to incorporate modeling features of the cognitive complexity of examinee responding behavior. They note three important points about EPM models: (1) it is unknown if there is any substantive psychological justification for these models at the level of item performance; (2) the models often make simplistic assumptions about the psychology of items; and (3) the “psychology of the test as a whole is left implicit” (p. 267) or is omitted entirely. Indeed, the a priori justification of the IRT model does not exist, and validation is carried out ex-post facto based on outcomes obtained by the model.

While IRT models suggest that knowing a given item’s parameters and an examinee’s θ are sufficient to predict a response, cognitive and learning theories may not support that assumption. In fact, Snow and Lohman note a number of studies refuting the IRT premise that simple proficiency is sufficient to predict all real examinees’ item responses. Differing strategies on tests (French, 1965) and on items within tests (e.g., Sternberg & Weil, 1980) impact item response behavior. Snow and Lohman assert that $\hat{\theta}$ is not indicative of a single latent proficiency; rather, it is a complex interaction of *different* types of knowledge, information processing, and strategies, some of which may vary across tests or persons. The authors further assert that components of EPM models such as item difficulty are not likely to be unidimensional. They state that cognitive science predicts many sources of item difficulty which are rarely if ever modeled. MIRT and other augmented or highly dimensional designs are attempts to reflect these complexities. Snow and Lohman elaborated further on the many cognitive correlates that are often ignored in educational measurement. Despite their warnings about the inaccuracy of overly simplistic EPM models, few published simulation studies utilize designs capturing this cognitive complexity. Consequently, measurement models such as

IRT may be unable to fully capture real-world testing behaviors. As such, simulations based on those models may fail to replicate real examinee responses. There are, however, some extant models in the literature that attempt to capture the complexity of test responses, such as the Linear Logistic Test Model (LLTM). The LLTM is based on a Rasch model linear combination of item properties, and allows for multiple item properties (e.g., de Boeck & Wilson, 2004). Within LLTM, some authors have extended the model to encompass many non-cognitive components, which may improve modeling of some real-world data (e.g., Kubinger, 2009).

2.3 Computer-Adaptive Testing Methods

While numerous formats for CAT exist, the purpose of this study is a basic analysis of score comparability for simulated and real linear P&P test data as utilized in a post-hoc CAT format. Post-hoc CAT formats use real examinee responses from real-world data in a CAT format as if those responses were the CAT-examination derived responses. This paper will examine the relatively simple item-level adaptive testing format, though many different formats exist (for an overview, see Wainer, 2000). To clarify the process of item selection and scoring, it may be instructive to review the basic functioning of an adaptive test. Simply stated, a traditional item-level adaptive test estimates θ after each item is answered by the examinee. The updated $\hat{\theta}$ determines which item is presented next by way of the information function. That is, the item with the highest information value at that particular proficiency estimate, is the next item chosen. This iterative process continues on until some predetermined stopping point is reached, such as test precision (i.e., standard error of $\hat{\theta}$) or number of items reached. A generalized list of steps is presented below to clarify the process when assembling an item level adaptive test with known item parameters (adapted from Wainer, 2000 and Mills & Stocking, 1996).

1. Starting Rule: Some reasonable rule should be in place for how to begin the adaptive test, since items cannot be chosen by proficiency estimate, because there is no $\hat{\theta}$ at the beginning of the test. Typically, items of moderate difficulty are chosen in a specified manner, such as random selection.
2. Item Selection: After the first few items are chosen, the examinee's responses to items of known parameters permits calculation of estimated, provisional $\hat{\theta}$. Based on the provisional $\hat{\theta}$, the next item is selected using some kind of criterion, such as maximum information for that $\hat{\theta}$. Given that maximum information is the target, an examinee's response will determine the next appropriate item, with different selections resulting for correct versus incorrect responses.
3. Test Completion: After selecting the next item, and the obtaining a response from the examinee, the next item is selected and the process continues until the test precision threshold is met, or some other stopping rule is applied. At the termination of the test, the final proficiency estimate is obtained using the complete set of item responses. A numerical score is typically assigned to reflect the proficiency estimate.

In IRT, item information functions provide a measure of a given item's contribution to proficiency estimation at a given point in the θ range. Some additional information about the relationships between item parameters and information may be instructive (as outlined in Wainer, 2000). Items with high pseudo-guessing (*c*) parameters contribute less item information and what they do contribute is maximized on the proficiency continuum just above the difficulty (*b*) parameter. Items with low discrimination (*a*) values are low on item information and do not contribute much to $\hat{\theta}$. While high discrimination items (*a*) are desirable, they contribute the most information only in a narrow range of $\hat{\theta}$. Information is the reciprocal of the squared standard error of estimation, *SE*($\hat{\theta}$). Information is a function of relating item parameters to proficiency estimates.

$$I_j(\hat{\theta}) = \frac{\left[P_j'(\hat{\theta})\right]^2}{\left\{P_j(\hat{\theta})\left[1 - P_j(\hat{\theta})\right]\right\}}$$

where $\hat{\theta}$ is the provisional proficiency estimate for examinee *i*; $P_j(\hat{\theta})$ is the probability of a correct response to item *j* given $\hat{\theta}$; and $P_j'(\hat{\theta})$ is the first derivative of $P_j(\hat{\theta})$ with respect to θ evaluated at $\hat{\theta}$ (both the formula and description are from Wainer, 2000).

2.3.1 Simulations with Post-Hoc CAT Designs

Attempts to improve modeling precision have been met with mixed success. In transitioning from traditional linear P&P testing to CAT designs, it is not uncommon to have a flurry of research activity designed to determine the comparability of a new CAT design to the P&P design. Often, researchers employ simulated data in their CAT systems analyses (Mills & Stocking, 1996); some researchers may be able to utilize their existing P&P test data in a post-hoc CAT simulation (e.g., Weiss, 2005a; Wang, Pan, & Harris, 1999). This method allows the researcher to determine the comparability of P&P outcomes to those of the new CAT design.

Wang, Pan, and Harris (1999) analyzed a post-hoc CAT design for Law School Admission Test data, comparing original $\hat{\theta}$ and CAT $\hat{\theta}$ as well as determining if a single administration was sufficient to obtain convergence at various precision levels. Using 969 actual examinees and 127 items (four sections with three major content domains) from the full test administration of the LSAT, they examined $\hat{\theta}$ recovery, and number of items necessary to complete a CAT at three levels of test precision, defined as standard error of measurement (SEM). Using the 3PL maximum likelihood proficiency estimation and maximum information method, they used 127 items actually administered to the examinees as the total item pool for their CAT. For the highest precision level, the authors found that 127 items were sufficient for estimating proficiencies for most, but not all, examinees. Moreover, using the real test data through a CAT simulation did not result in adequate θ recovery, except in the highest precision method. The authors conclude that the small item pool negatively impacted their results. Unfortunately, comprehensive cumulative data were missing, such as additional in-depth analyses and item information summaries. The inclusion of all four LSAT operational sections on one test may also have made some results unclear given unmodeled multidimensionality effects. Without CAT replications and no indication of programming specifics used to create the CAT, the reader is left without sufficient information for a critical review.

One approach to improve simulation modeling is to add dimensions to the 3PLM without regard to any a priori categories. This multidimensional approach can recover the complexity of real data more accurately than the simpler 3PLM. A similar approach is to add a priori categories to augment the 3PL model. This approach has the benefit of conceptual categories for added parameters. Utilizing real examinee responses in a CAT simulation design addresses the limitations of more basic simulations and facilitates understanding of the comparability of outcomes when using real data in a CAT.

Highly multidimensional modeling of real data may improve recovery of examinee responses, but it is an atheoretical position which may prove problematic when used on different types of data. The problem of overfitting remains, and it may be a sufficiently significant limitation to diminish generalizability of this approach. Adding only two new components, the a priori augmented modeling approach provides somewhat improved recovery of the original data structures; in most cases the improvements from the simpler 3PL model were relatively trivial. These attempts to use real data to improve modeling for simulations did not address the use of real data directly within a CAT design. Unfortunately, the post-hoc approach failed to utilize a rigorous methodology and failed to analyze the finer level differences between real and simulated data.

Harwell et al. (1996) note the ubiquity and influence of measurement based simulations, yet they also note the failure of researchers to use rigorous scientific methods. These limitations can lead the reader to doubt that simulations accurately represent real-world test taker behavior. What is missing from the literature is a rigorous post-hoc CAT study comparing real examinee responses to simulated responses using sensitive statistical analyses such as RMSE to determine the extent to which simulation designs can recover important aspects of real examinee response behaviors.

Chapter III. RESEARCH METHODS

The research design seeks to evaluate the equivalency of proficiency estimates derived from simulated data to those derived from real data. The basic design of the study includes two data types for comparison, real and simulated. These datasets were utilized in post-hoc item-level CAT design as well as in creating artificial linear (P&P) tests, with three variable test lengths. The results were analyzed using both broad and fine comparisons of $\hat{\theta}$ including bias, RMSE, and Pearson correlation. In addition, summaries of classical test theory statistics for linear tests and item exposure rates for CAT designs were obtained.

3.1 Data Considerations

3.1.1 Real P&P LSAT Data

Real examinee data were obtained from 20 administrations of the LSAT, a large-scale standardized, linear, paper-and-pencil professional school admission examination. Data from the LSAT administrations were extracted only for the three Logical Reasoning (LR) item sets composed of approximately 50 scored and 25 unscored multiple choice items per administration. The LR administrations were chosen because they are highly unidimensional, and a fundamental assumption of IRT is trait unidimensionality (Hambleton, Swaminathan, & Rogers, 1991).

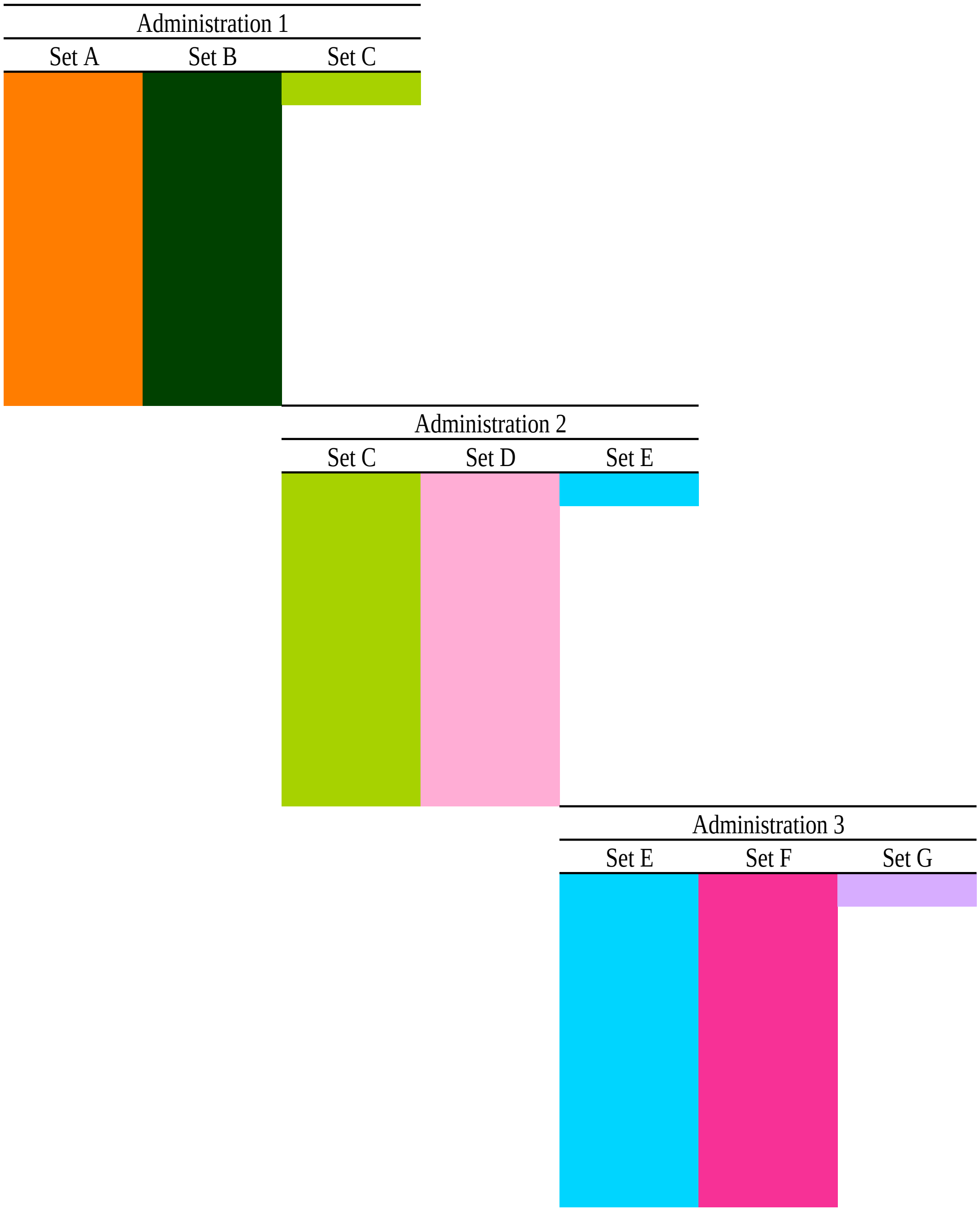
3.1.2 Creating the Synthetic Examinee

To obtain a true estimate of proficiency as well as a CAT item pool requires a large number of items with real responses. In this study, a large vector of item responses was required, but unavailable. Therefore, it had to be created. Searches of the literature did not locate any information on methods for creating a large item response dataset with synthesized response vectors across test administrations. Therefore, a method was developed for this study. The goal was to maximize logical, reasonable choices while minimizing potential limitations. To accomplish this task, the following key points were addressed: the structure of the examination, the matching of response vectors, the completion of the full matrix, and the matching of the synthetic examinee across administrations.

3.1.2.1 Structure of the Examination Data

The first consideration in creating the synthetic examinee is to clarify the data structure of the examination administrations. All scored data were obtained in two sets of approximately 25 scored items. This combined 50 item set was administered to all examinees numbering from approximately 20,000 to more than 40,000 per administration. During each test administration, one additional set of approximately 25 non-scored items is administered to a subgroup of around 2,000 examinees. Therefore, all examinees respond to about 50 items, and a small percentage respond to an additional 25 items which are not scored. These non-scored items are re-administered as a scored section in a subsequent test administration. Thus, for a subset of examinees on a given test administration, approximately one-third of all items will be shared in common with a different administration. To illustrate the structure graphically, it may be helpful to review Figure 1.1. The column size (i.e., length) represents the number of examinees and the colors represent items sets. Blocks of the same color use the same items.

Figure 3.1 Example of the Structure of the LSAT Logical Reasoning Sections



3.1.2.2 Matching Item Response Vectors

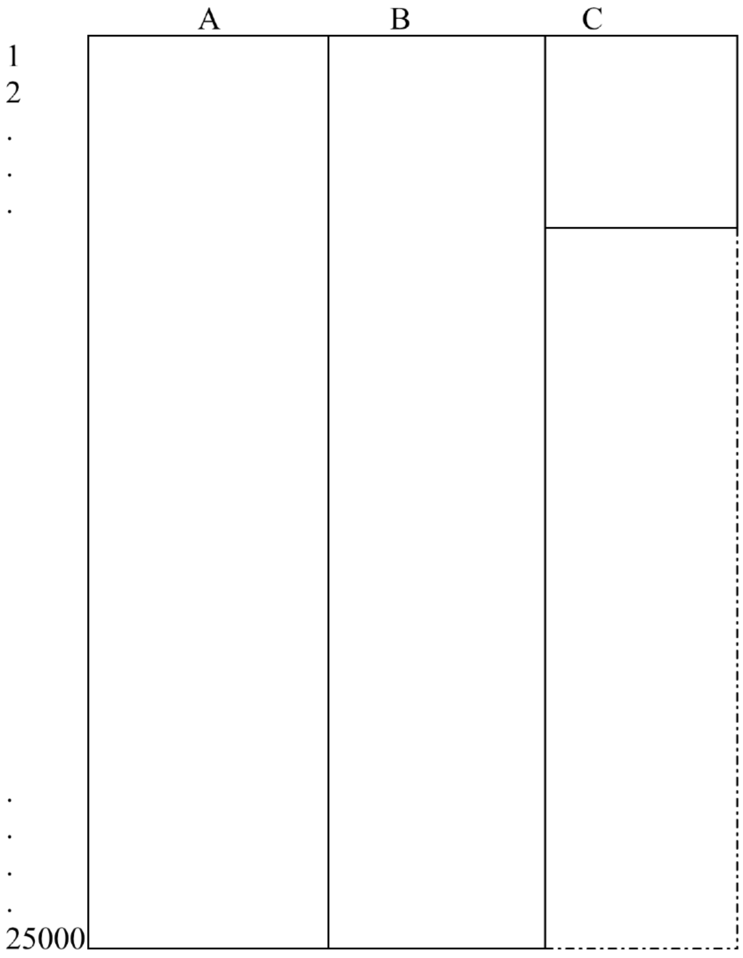
To match item responses vectors, it is necessary to obtain identical or nearly identical response strings. While computer science and information theory provide some compelling methods called distance measures, a pilot study for this dissertation demonstrated inaccurate matching when analyzing simple binary vectors which have important position dependencies. Instead of using these complicated methods, a simple overlap measure proved to be more effective at matching response strings. Binary response vectors were compared by creating an overlap variable defined as the number of identical item responses located in the same item positions. A simple example will illustrate the concept for response vectors of equal length.

Let A = 1110111 and B = 1**0**101**0**1. The overlap measure between the strings is 5 because 5 is the number of items matching in the same positions. SAS software was used to calculate this variable for every response vector comparison.

3.1.2.3 Creating the Complete Response Matrix

If an examinee's response pattern is nearly identical on a given set of items within an administration, then those examinees can be considered very similar in terms of proficiency as well. The 3PL model requires moderately large numbers of examinees to facilitate stability of item parameter estimates. Using these datasets in their original formats, it is impossible to utilize more examinees than are available in the non-scored linking section. It is possible, however, to replicate the examinees available so as to complete the matrix as shown in Figure 2. A and B are 25-item scored response vectors and C is the 25-item non-scored response vector administered only to the smaller subset of examinees. The dotted line represents the replicated examinees from the non-scored items section.

Figure 3.2. Graphical Illustration of a Single Administration Dataset



Replicated response vectors were added to the administration datasets by obtaining matches on item sets A and B for examinees with C responses. For examinees with item set C responses, their A and B vectors will be matched with examinees with no C responses using the maximum response vector overlap approach. If several examinees had patterns with the same overlap score, the following additional matching criteria were used in the designated order: gender, ethnicity, and age. If more than one examinee matched on response patterns as well as on the additional demographic criteria, the final matching vector was selected at random from among the extant perfect matches. After obtaining the final matching vector, the section C item responses were copied to the matching A+B-only examinees, creating the full 75-item response set. Therefore, response strings from the smaller set of items were used more than once to create the full matrices for each administration. The best matching vectors were retained and added to the original full set of response vectors. The resulting final matrix was set to be approximately 75 items by exactly 12,000 examinees, obtained by retaining the best 12,000 matches. Because the first dataset was the baseline dataset, it was reduced to its final number of 5,000 examinees immediately using random selection from among the best 12,000 examinees.

3.1.2.4 Matching Between Administrations

Given the structure of the data, an ideal method of linking administrations (created as described in section 3.1.2.3) is by using the common items overlapping between them. Matching using IRT-based methods would confound the analysis of IRT outcomes later. Therefore, the goal was to create a method of matching response vectors between administrations without using IRT-based approaches. Unfortunately, no such matching procedure was found in the literature or in practice. Any researcher attempting to link datasets in this manner will have any number of options, but the researcher will likely need to keep computing time to a reasonable level; response vector matching across large datasets can use considerable computing resources. Given that limitation, the goal was to create the best possible matches so that the resulting synthetic examinee will be as realistic as possible.

After the full data matrices were created, synthetic examinees were matched between administrations using the common-item response vectors. The between-administration matching was accomplished by using the maximum overlap approach as described previously (section 3.1.2.2). A number correct score for overlapping items was calculated for every examinee. Examinee vectors were compared to vectors with the same number correct values, if possible. This step was added to reduce the computational burden of comparing strings which are unlikely pattern matches because they have differing number correct scores. Synthetic examinees' non-scored vectors were randomly selected from the current administration dataset and matched with the scored response vector of the same items on the linking administration. Failing to select the order of comparison randomly could result in unacceptably poor matches for the vectors at the end of the data file for which there would be far fewer match options.

If the number of examinees available for comparison was greater than or equal to 500, then matching was made directly from within that group. If there were fewer than 500 examinees available for comparison, then the pool of available comparison vectors was expanded to include vectors with number correct scores plus or minus one (±1) from the reference examinee. If the newly expanded pool again failed to have 500 or more vectors available for comparison, then the pool was expanded again to include number correct scores of ±1 and ±2 the reference number correct. Only vectors with the best overlap scores were retained for comparison. If there was more than one matching vector with the same overlap value, demographic factors were used to refine the match. The demographic factors used to augment the matching were the following: gender, ethnicity, and age. If more than one match existed after these three additional data points were added, a random selection procedure was used to choose the final match from among the perfect matches. So that each examinee may be chosen only once, previously matched vectors were unavailable for subsequent matching. This process continued until all examinees were matched.

After completing the matching, the replicated item set was dropped from the dataset, retaining the best 5,000 cases. These 5,000 examinees were chosen using the best matches from the within-administration matching procedure. The final product of the first matching process, therefore, had 100 items and 5,000 examinees. These 5,000 examinees were subsequently matched to the next matching synthesized 12,000-examinee dataset. Matching the 5,000 to 12,000 avoids the problem of having to match to more poorly matching vectors which would be the result if each vector had only a one-to-one match option. After all matching between datasets was completed, 550 items were obtained with 5,000 synthetic examinees. This dataset was retained as the real, synthesized dataset.

3.1.3 Data Calibration and Proficiency Estimation

BILOG-MG 3.0 for Windows® (Zimowski, Muraki, Mislevy, & Bock, 2003) was used to obtain proficiency estimates and item parameters from the synthesized dataset. The program option for the 3PL model was used, because it is a commonly used model in educational measurement. Fixing the item parameters, the examinees' proficiencies were estimated using Bayesian expected a posteriori (EAP) methods. Because of the large number of items utilized, the proficiency estimates are considered to be reliable indicators of the examinees' true proficiencies. Sixteen items failed to converge and were dropped from the analysis to facilitate convergence and adequate fit.

3.1.4 The Simulated Data

As stated previously, simulation designs utilize data created by using randomly drawn variables from specified distributions. A set of typical steps were taken to simulate item response data. The sample size of 5,000 was set to mimic our real data set. The simulee values for θ were used from the proficiency distribution obtained from scoring the real data. Using the real θ ensures that simulated data most closely conform to the real data specifications.

Harwell et al. (1996) state that randomly generated item parameters are particularly unrealistic, often resulting in combined *a*, *b*, and *c* item parameters which are unlikely in real item calibrations. Therefore, item parameters from the calibration output of the real data were utilized. Substituting the true proficiency values and item parameters into the 3PL equation creates a matrix containing the probabilities for simulee *i* on item *j* (represented as p_{ij}). Therefore, the generated matrix was of dimensions 5000 x 534. From this matrix of probabilities based on $\hat{\theta}$ and item parameters, the matrix of binary scored response values was created using a random univariate distribution for comparison. Using the SAS random univariate (RANUNI) option, a univariate data distribution was created with values in the range (0,1). As described by Fan, Felsovalyi, Sivo, and Keenan (2001), the SAS RANUNI option utilizes a congruential generator which Harwell et al. (1996) describe as desirable for simulations due to their facility at producing sufficiently random data for simulation accuracy. To complete the data simulation matrix, a value from the univariate distribution (u_{ij}) was randomly selected and compared to the model-derived p_{ij} matrix value. If the selected value was less than or equal to the value of p_{ij} then the item was scored as correct (indicated by a “1”). Random values greater than p_{ij} were scored as incorrect (indicated by a "0"). Once the data matrix was created in this manner, it was used as one simulated dataset.

To reduce potential sample bias in simulations, it is common to create many replication datasets. Harwell et al. (1996) recommend that replications reflect the design used for the study, increasing or decreasing based on the demands of the study variables. For most IRT studies, Harwell et al. recommend a minimum of 25 replications, though they admit that the number could be much higher for some studies. Statistical power is also affected by the number of replications, with too few providing insufficient power. To ensure the smallest amount of sampling bias, to obtain adequate power, and to satisfy the requirements of inferential analyses, the total number of replicated datasets created was 100.

3.2 Number of Items

The number of items used in any examination is often a compromise between examinee burden and proficiency estimation efficiency and precision. Computer-adaptive tests can target items for maximal information at a given examinee's proficiency level, but linear P&P tests are unable to utilize such efficiencies. For both CAT and P&P tests, the optimal number of items is typically the minimum number necessary to obtain stable proficiency estimates for all examinees while keeping within time and other constraints. The current LSAT P&P examination uses 50 scored LR items. To determine the optimal number of items for CAT and linear tests using both real and simulated data, item numbers were manipulated based on the P&P test length such that 50%, 70%, and 100% length tests were examined. As such, the number of items examined was as follows: 25, 35, and 50.

3.3 Artificial Computer-Adaptive Test

Neither the simulated nor real data were obtained from computer-adaptive test administrations. The real data were taken from the LSAT, a paper-and-pencil linear test; the simulated data are generated data, never administered to any examinees in reality. To determine the comparability of $\hat{\theta}$ from the original datasets in a CAT requires that both types of datasets be used to simulate an examinee's (or simulee's) progress through the CAT. Given that the simulated datasets were replicated 100 times, both the simulated and real data were used in a CAT exam format 100 times. To facilitate this step, an existing software application, POSTSIM 2.0, from Assessment Systems Corporation (Weiss, 2005b) was utilized. The program output includes detailed CAT simulation information including $\hat{\theta}$ for each examinee/simulee. Both real and simulated data were utilized in the post-hoc CAT format examinations.

To start the CAT, examinees were assumed to have no proficiency estimates, so the CAT software was set to use some simple method to select initial items. For this CAT, five items were randomly selected from among the 200 items with the highest information values. Responses to these items allowed the program to calculate a provisional $\hat{\theta}$ so that traditional maximum information item selection procedures could be used to choose subsequent items to administer. Proficiency estimates were obtained using the Bayesian EAP method. The CAT stopping rule was defined by the three test lengths noted previously. Records of items used were retained to provide item exposure summaries.

3.4 Artificial P&P Linear Test for Simulated Data

Unlike $\hat{\theta}$-targeted CAT item selection, a linear test must use items with a broad spectrum of item parameters in the hope that they will provide sufficient information for the full distribution of examinee proficiency. Therefore, the simulated linear tests used items selected from a stratified sample of items which were ordered by traditional proportion correct (*p*-plus) difficulty values. Depending on the test length manipulation, a suitable number of equally-spaced strata were created and used to randomly select items from within each stratum. One item from each stratum was randomly chosen and that item was included on the linear test until the test length specification was reached. Once artificial linear tests were created, the items used in real datasets were used to create linear tests for the 100 simulated datasets. Using the same linear items will facilitate direct comparison of results between real and simulated data. Since there are 100 simulated datasets and three test lengths, 300 simulation data linear tests were created.

3.5 Manipulations and Outcomes

CAT and linear testing formats, real and simulated data types, and three test lengths were manipulated to illustrate differences to the main outcome measure, estimated proficiency. Additional outcome variables such as proportion correct, biserial correlations, and relative test efficiency were included to augment the proficiency summaries. Specific manipulations are described in the following section.

3.6 Data Types and Scoring Method

A comparison between CAT and linear forms of both types of data were completed to show summaries of proficiency estimates and item exposure rates. Proficiency estimates were obtained using EAP methods. Item exposures will be summarized to show how often items are administered to examinees.

3.7 Analyses and Comparisons

Analyses were of two types, broad and fine. Broader analysis included proficiency means as well as correlations. For linear tests, classical item difficulty (*p*-plus), discrimination (biserial correlations), and $\hat{\theta}$ will be summarized with means and standard deviations as well as Pearson correlations for all conditions.

Finer level analyses examined θ estimation recovery for simulated data as compared to the real data. Both bias and root mean squared error (RMSE) measure differences between estimated and true parameter values. Bias is represented by the following formula:

$$\text{Bias}=\frac{\sum_{i=1}^{N}(\hat{\theta}_i-\theta_i)}{N}$$

where $\hat{\theta}_i$ is the estimated proficiency for simulee *i*; θ is the true proficiency parameter derived from the full data matrix; *i* is the simulee index; and *N* is the total number of simulees.

The following equation represents the RMSE formula for comparing the real θ to $\hat{\theta}$ from replicated sampling:

$$\text{RMSE}=\sum_{j}^{K}\left(\sqrt{\frac{\sum_{i=1}^{N}\left(\hat{\theta}_i-\theta\right)^2}{N}}\right)/K$$

where *j*=replication index and *K*=total number of replications.

The standard error, $SE(\hat{\theta})$, statistic was calculated as listed below.

$$\text{Standard Error } (\hat{\theta}) = \sqrt{\frac{1}{I(\hat{\theta})}}$$

Detailed information will be helpful in illustrating any potential differences: Analyses included calculations for standard error, bias, RMSE, and Pearson correlations. Examinee-level factors were analyzed for each simulee by calculating the empirical standard deviation of the $\hat{\theta}$ across replications and comparing this value to the calculated mean standard error statistic. This analysis illustrates the accuracy of the mean standard error statistic versus an empirical standard deviation of $\hat{\theta}$ across replications.

Finally, relative efficiency is a measure of the efficiency with which a given test measures proficiency. It is calculated using the following formula which utilizes the $I(\theta)$ function noted previously.

$$RE(\theta) = \frac{I_A(\theta)}{I_B(\theta)},$$

where $RE(\theta)$ denotes relative efficiency $I_A(\theta)$ and $I_B(\theta)$ are the information functions for tests A and B defined over the common proficiency scale, θ (Hambleton, 1993). The interpretation of the relative efficiency value would be such that a value of 2.0 would denote that test A has twice the cumulative information as test B; a value of 1.0 would denote that test A and test B have the same cumulative information; and a value of less than 1.0 would denote that test B has more cumulative information than does test A.

CHAPTER IV. RESULTS

In this study, 904 different datasets were created and analyzed for both real and simulated data, within linear and computer-adaptive administration formats, and with three test lengths (see Table 4.1 for summary). Each of these three test lengths comprised the following datasets: a linear test using the real data, a linear test using the simulated data, a CAT test using the real data, and a CAT test using the simulated data. For this paper, the term "real data" refers to the data from the artificially synthesized dataset which was created from the true LSAT data as described in the previous methodology section. The real data were used as they are, from the one synthesized real LSAT dataset. For simulated data, all datasets were simulated using the parameters taken from the synthesized dataset. Linear tests used the items selected in advance, whereas CAT tests were assembled real-time from the pool of 534 items. Each of the 100 CAT tests used one of the 100 simulated datasets, resulting in 100 total simulated data computer-adaptive tests. The one real dataset was used 100 times by CAT, with each iteration choosing items as dictated by the maximum information algorithm. Additionally, the real dataset of 534 items was analyzed.

Table 4.1

Summary of Datasets Used in Analyses

Dataset	Number of Test Items	Iterations of Each Test	Total Datasets
Full Synthetic Dataset	534	1	1
Linear Format-Real Data	25, 35, 50	1	3
Linear Format-Simulated Data	25, 35, 50	100	300
CAT Format-Real Data	25, 35, 50	100	300
CAT Format-Simulated Data	25, 35, 50	100	300
TOTAL			904

4.1 Complete Synthetic Examinee Data

As described in the methodology section, the complete matrix of synthetic examinees was created from the original examination data. The final number of items was 534 and the final number of examinees was 5,000. The resulting unique number correct raw scores and proficiency estimates are summarized in Table 4.2. The data were analyzed by sorting the data from lowest proficiency estimate to highest and assigning each group of 500 ordered cases a group value for a total of 10 groups. Additional summaries throughout this study also utilized this grouping method. With 87 unique proficiency estimates, Group 1 had the largest number of unique values. Group 10 had the second largest number of unique proficiency values at 41. Groups 1, 2, 3, 5, 9, and 10 were found to have 15 or more unique proficiency estimates, whereas Groups 4, 6, 7, and 8 had fewer than 10 unique values. Group 4 had only 4 unique values, and Group 8 had only 5, giving these two groups the smallest number of unique proficiency estimates. These unique proficiency values may impact the analyses of results in this study, as within-group analyses with fewer values will result in less variation. Smaller proficiency variance values result in smaller summary statistics such as standard deviation. For the additional summary of the complete matrix data, the following statistics for the full synthetic matrix are reported: Classical proportion correct (i.e., p-plus) and biserial correlation for linear test data, and IRT descriptive statistics for proficiency estimates, item parameters, and item information.

4.1.1 Complete Synthetic Matrix Classical Test Indices

Table 4.3 presents a summary of the mean, standard deviations, and median of examinee-level *p*-plus for the complete synthetic matrix. In addition, the overall biserial correlation was computed but will be reported only in the text. As expected, the examinee p-plus values increased as examinee proficiency increased. In Group 1, the lowest proficiency group, the mean *p*-plus value was 0.286. In Group 10, the highest proficiency group, the mean *p*-plus was 0.944, an increase of 0.658. This outcome is expected for all test groups in this study because it is reasonable that more able examinees would mark a larger proportion of items correctly. Within the proficiency groups, the standard deviations were found as expected, (i.e., with greater score variability in the tails of the score distribution and less variability in the center). The overall *p*-plus mean of 0.681 indicates moderate item difficulty (Crocker & Algina, 1986). The item-level biserial correlation, which is a classical measure of item discrimination, was calculated to have a mean value of 0.562 with a standard deviation of 0.190 and a median biserial value of 0.566. These values indicate an appropriately

Table 4.2

Complete Matrix Summary of Unique Proficiency Estimates and Number Correct

Group	Unique Raw Score Values	Unique $\hat{\theta}$ Values
1	65	87
2	29	20
3	23	17
4	18	4
5	14	16
6	9	8
7	16	7
8	8	5
9	13	17
10	29	41
ALL	224	222

moderate discrimination levels for this set of items (Crocker & Algina). Group-level calculations were not completed for any correlation indices because of the problem of range restriction which creates inaccurate correlation results (e.g., Bobko, 1983; Lord & Novick, 1968). Groups 1, 2, and 10 showed the largest standard deviation (SD) values for proportion correct, indicating greater variation in scores for these groups.

Table 4.3

Full Synthetic Dataset Proportion Correct Statistics by Group

Group	Mean	SD	Minimum	Median	Maximum
1	0.286	0.076	0.078	0.288	0.420
2	0.470	0.034	0.404	0.480	0.512
3	0.570	0.022	0.512	0.570	0.616
4	0.636	0.021	0.606	0.636	0.676
5	0.692	0.017	0.664	0.698	0.716
6	0.728	0.015	0.712	0.728	0.750
7	0.777	0.015	0.750	0.778	0.818
8	0.833	0.014	0.818	0.826	0.866
9	0.874	0.018	0.832	0.876	0.896
10	0.944	0.025	0.896	0.940	1.000
ALL	0.681	0.191	0.078	0.712	1.000

4.1.2 Linear Test Classical Indices

As with the full synthetic dataset, classical item level *p*-plus and biserial correlation indices were calculated for the three linear format tests. The item level p-plus summaries are presented in Table 4.4, Table 4.5, and Table 4.6. As expected, mean p-plus values increased as proficiency increased and the standard deviation is larger towards the middle groups and smallest at the extremes. The overall mean *p*-plus increased slightly as test length increased from 25 items (0.631) to 35 items (0.637) to 50 items (0.638). The overall *p*-plus SD was 0.185 for the 25-item test, and a similar but smaller 0.178 for the 50-item test. The 35-item test *p*-plus SD, however, was 0.192 which was the largest overall SD of the three test lengths. Similarly, the biserial correlations for the 25- and 50-item tests were quite similar at 0.615 and 0.616 respectively, while the 35-item biserial dropped to 0.592. In this case, the 35-item test demonstrated reduced discrimination ability which can negatively impact the stability of proficiency estimation procedures. These results indicate that the linear tests performed basically as expected with some deviation from expected patterns shown for the 35-item test.

Table 4.4

Linear Test Real Data Classical Indices – 25 Items

Group	P-plus Mean	SD of P-plus
1	0.237	0.164
2	0.397	0.227
3	0.489	0.237
4	0.566	0.237
5	0.629	0.235
6	0.693	0.224
7	0.744	0.211
8	0.800	0.186
9	0.858	0.156
10	0.931	0.103
ALL	0.631	0.185

Table 4.5

Linear Test Real Data Classical Indices – 35 Items

Group	P-plus Mean	SD of P-plus
1	0.252	0.167
2	0.420	0.219
3	0.509	0.229
4	0.576	0.230
5	0.635	0.229
6	0.685	0.228
7	0.730	0.222
8	0.779	0.206
9	0.827	0.190
10	0.893	0.158
ALL	0.637	0.192

Table 4.6

Linear Test Real Data Classical Indices – 50 Items

Group	P-plus Mean	SD of P-plus
1	0.233	0.140
2	0.406	0.196
3	0.494	0.211
4	0.566	0.217
5	0.625	0.219
6	0.678	0.217
7	0.724	0.211
8	0.771	0.196
9	0.820	0.175
10	0.888	0.129
ALL	0.638	0.178

4.1.3 Full Matrix IRT Results

IRT results include calculations of proficiencies, standard errors, item parameters, bias, and RMSD. Table 4.7 summarizes the results of the IRT proficiency values by group. The expected mean for a proficiency distribution is 0.0 and the expected standard deviation is 1.0. The table shows that, for the full synthetic examinee dataset, the overall mean is -0.003, and the standard deviation is 1.014. These values are similar but not identical to the expected values of 0 and 1. At 0.695, the group with the highest score standard deviation is Group 1, the lowest proficiency group. The SD for Group 1 is several times the size of the next largest SD and indicates a large variation in scores within that group. The smallest SD was found within Group 4 and Group 8, both found to have a very low SD value of less than 0.001. Both of these groups are the moderately low and moderately high ability groups and the proficiency score homogeneity within these groups is very high (as noted previously in Table 4.2). Similarly, Group 4 had the smallest standard error value at 0.002, and Group 8 had the second smallest SE value of 0.006. Both groups had notably smaller values than all other groups.

Table 4.7

Full Synthetic Matrix Proficiency Estimate Statistics (n=5,000)

Group	$\bar{\hat{\theta}}$	$SD\left(\hat{\theta}\right)$	$\overline{SE\left(\hat{\theta}\right)}$
1	-1.940	0.695	0.113
2	-0.884	0.214	0.034
3	-0.707	0.088	0.046
4	-0.247	0.000	0.002
5	-0.167	0.122	0.092
6	0.233	0.032	0.048
7	0.283	0.124	0.014
8	0.741	0.000	0.006
9	0.981	0.224	0.078
10	1.680	0.228	0.187
ALL	-0.003	1.014	0.062

The largest SE value of 0.187 was found in Group 10, and Group 1 had the second largest SE value of 0.113. In the remaining groups, the values ranged from 0.034 to 0.092. The overall mean SE was calculated to be 0.062. For Groups 2 and 8, the small mean standard error results indicate that proficiency estimates in these groups have less error associated with the values than would higher SE values, such as those found in Group 1 and Group 10. The latter two groups have proficiency estimates that have been calculated with a higher level of error.

Item parameter mean and standard deviation values for a test are an indication of item difficulty (*a*), discrimination (*b*), and pseudo-guessing (*c*). In Table 4.8, the overall full synthetic matrix item parameter values are summarized, along with the values for the three linear subtest forms. For the full synthetic matrix, the mean *a*-parameter value of 0.840 indicates a moderately high degree of item discrimination; the *b*-parameter mean is -0.506 and are, therefore, of moderate difficulty overall; the mean *c*-parameter value of 0.056 indicates that pseudo-guessing is fairly low (see Crocker & Algina, 1986). The pseudo-guessing parameter outcome can be interpreted as follows: On average, approximately 5.6% of low-ability examinees will answer an item correctly by guessing.

The 25-item test had a mean *a*-parameter value of 0.940 which indicates a fairly high degree of discrimination. On the same test, the mean *b*-parameter was -0.582, indicating that items were moderately easy. The mean *c*-parameter was 0.061 indicating that approximately 6.1% of low-ability examinees will answer correctly by guessing. The 35-item test had a mean *a*-parameter of 0.815, indicating a fairly high level of discrimination. The mean *b*-parameter for the 35-item test was -0.394 which can be interpreted as moderately easy. Among the three shorter test lengths, this *b*-parameter value is the closest to zero and is the value most different from the other means. The *c*-parameter mean for this test was 0.049, a slightly smaller value than was found for the 25-item test. The 50-item test had a mean *a*-parameter of 0.816 indicating a fairly high level of discrimination. The mean *b*-parameter for the 50-item test was -0.532 which indicates that items were moderately easy. The 50-item mean *c*-parameter was found to be 0.036.

Table 4.8

Linear Test Item Parameter Descriptive Statistics by Test Length

Test	$\bar{a}$	$SD(a)$	$\bar{b}$	$SD(b)$	$\bar{c}$	$SD(c)$
25	0.940	0.651	-0.582	1.078	0.061	0.097
35	0.815	0.611	-0.394	1.973	0.049	0.084
50	0.816	0.570	-0.532	1.265	0.036	0.061
534	0.840	0.542	-0.506	1.797	0.056	0.088

Comparing the three tests, the *a*-parameter standard deviation was highest for the 25-item test at 0.651, and lowest for the full 534-item test, at 0.542. The longer the test length, the more the *a*-parameter standard deviation decreased. Unlike the *a*-parameter patterns, the *b*-parameter standard deviation was largest for the 35-item test at 1.973. The second largest *b*-parameter standard deviation of 1.797 was found for the 534-item full test. The smallest *b*-parameter standard deviations resulted from the 50-item (1.265) and 25-item (1.078) tests. These results indicate that, after taking test length into account, the three test lengths have different characteristics. A potentially important reason for the somewhat erratic item parameter patterns is the imperfect item selection procedure. The three linear test length item sets were created using a random within-bin selection procedure which resulted in tests which were dissimilar in terms of item performance. The three real data linear tests were dissimilar to each other and dissimilar to the full matrix dataset in terms of item parameter means and standard deviations. As such, any analyses of differences between these short tests should be made with that understanding in mind. In particular, the 35-item dataset demonstrated a *b*-parameter mean and standard deviation that is notably different from the other two short tests as well as the full matrix dataset. Thus, comparisons between the three short tests should be made with caution.

Test Information. The adjusted test information presented in Figure 4.1 for the 25-item test length and Figure 4.2 for the 35-item test length, and Figure 4.3 for the 50-item test length, is the test information divided by the number of items. This adjustment puts the test information on the same scale and, therefore, provides a more direct information comparison across test lengths. Figure 4.1 illustrates the generally stable information function, though with some variability particularly in the lower proficiency range. As expected, information is higher towards the middle proficiency range and lower towards the tails or extreme proficiency values. Generally, most of the values seem to fall between 0.3 and 0.5 with some outliers, and the aforementioned decline in the tails. More information variability is apparent in the lower proficiency tail of the distribution. In addition, the adjusted means were calculated with the result that the 25-item linear test adjusted mean was 0.330 with a standard deviation of 0.111.

For the 35-item test, Figure 4.2 illustrates the results which indicate that most of the adjusted information values range between 0.2 and 0.4 with the predictable decline in the tails. The rightmost higher proficiency tail lost information more gradually compared to the lower proficiency tail which lost a large amount of information over a small area. Near the middle of the distribution there was a spike of information for a narrow band of proficiency, and high variability can be seen in the low proficiency tail. Within the lower proficiency tail, there are some higher spikes in information, which would not generally be an expected outcome. The 35-item adjusted mean information was 0.270 with a standard deviation of 0.104, a mean which is smaller than was found for the shorter 25-item test.

Figure 4.1 Information for 25-Item Linear Test Adjusted for Test Length

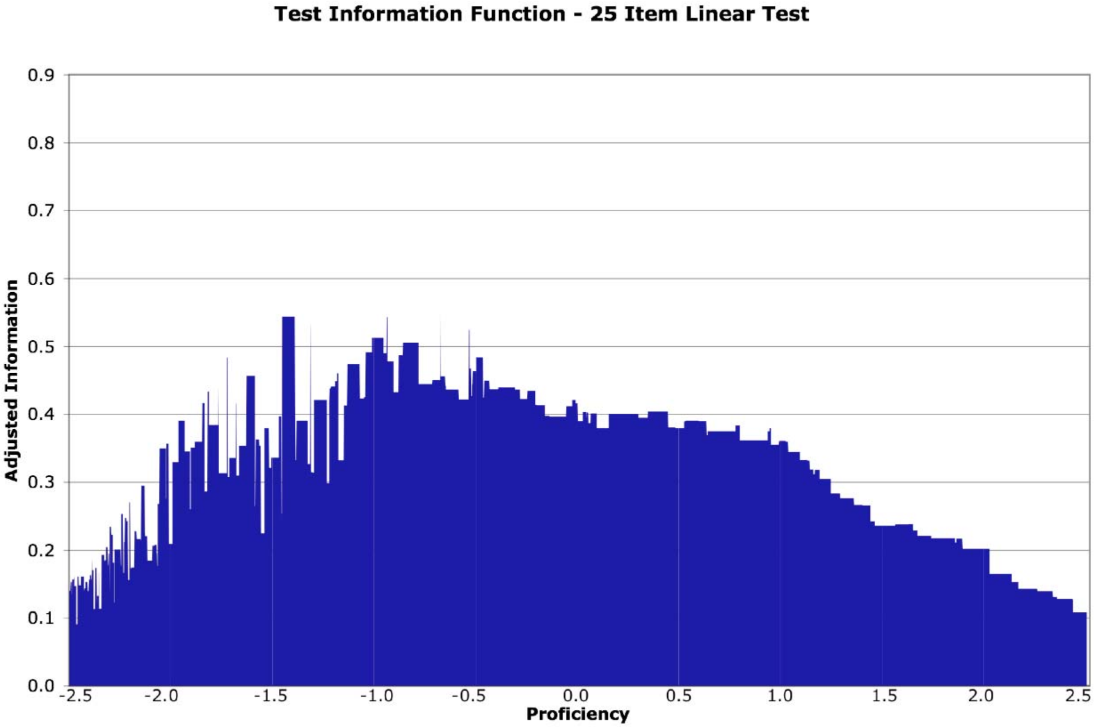
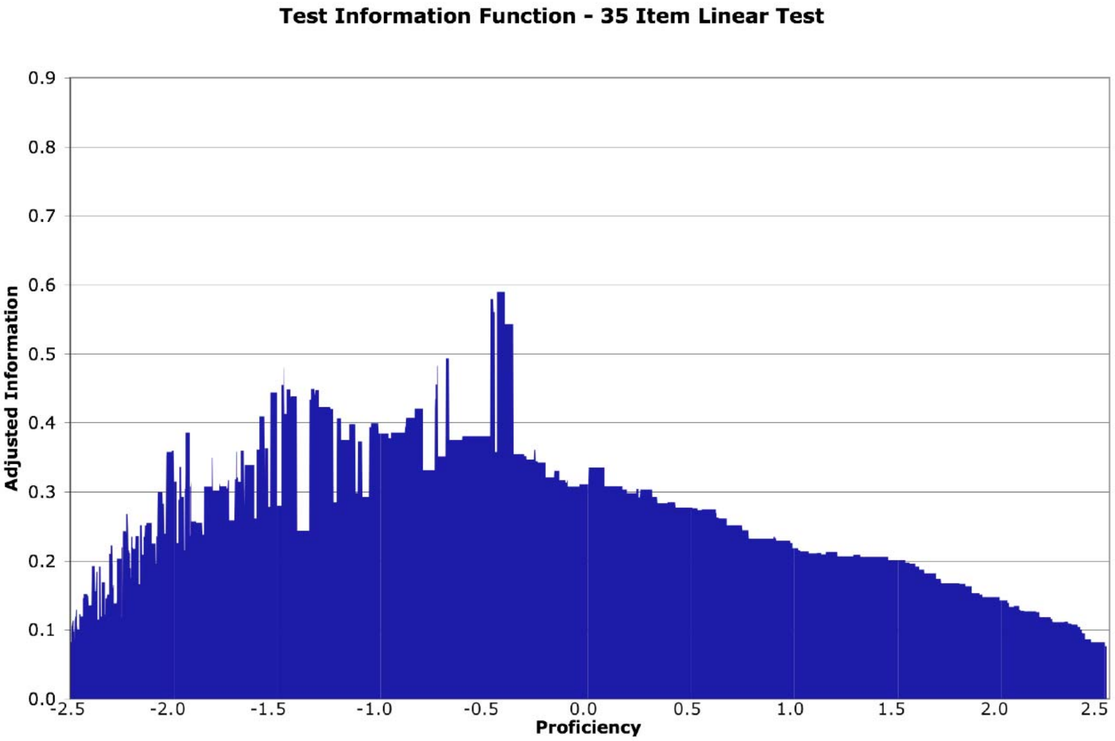
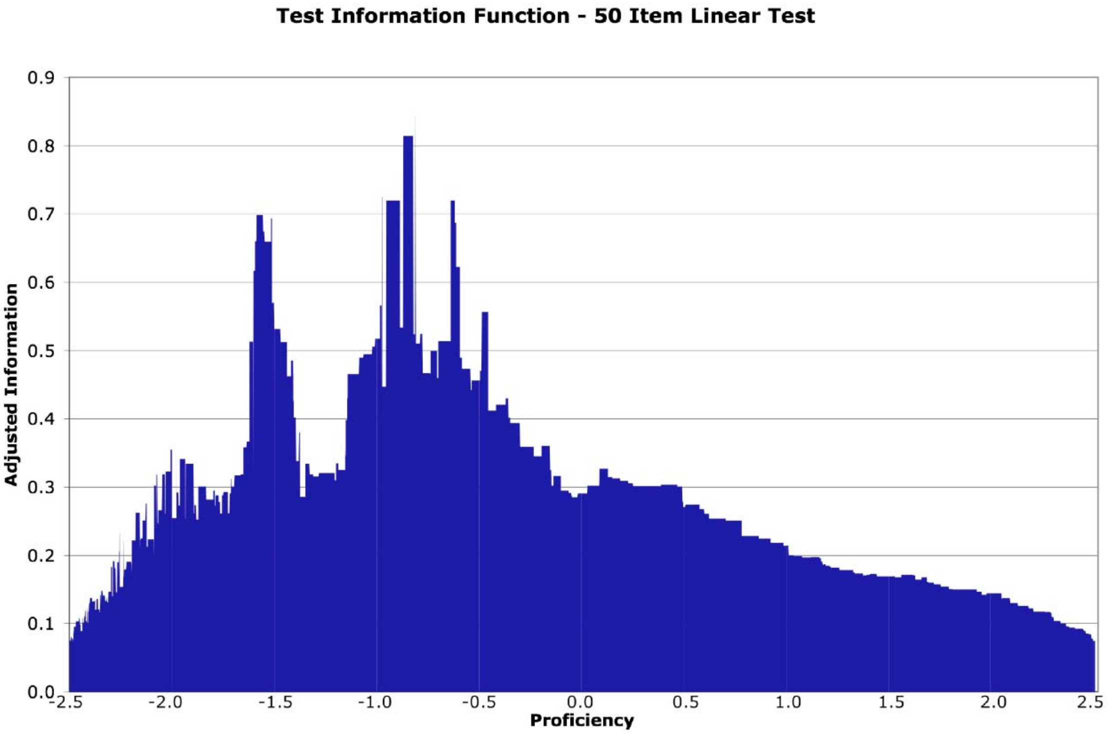


Figure 4.2 Information for 35-Item Linear Test Adjusted for Test Length



The 50-item linear test is summarized in Figure 4.3. Most of the values appear to fall between 0.3 and 0.5, with a few narrow peaks reaching 0.7 or above. The distribution does not appear to be a normal distribution. As with the other test lengths, the higher proficiency information gradually declines, compared to the sharp decline in the lower proficiency tail. The overall adjusted mean information was calculated to be 0.292 and the standard deviation was 0.154. The adjusted overall mean value is the middle value of the three test lengths, with the 25-item test having a larger mean and the 35-item test having a smaller mean.

Figure 4.3 Information for 50-Item Linear Test Adjusted for Test Length



4.2 Comparisons Across Administration Format and Data Types

Comparing real and simulated data conditions are an important focus of this study. Specifically, it is important to compare linear and CAT formats using real data and simulated data in each, and by using each test length to obtain and compare proficiency estimates, bias, RMSD, item information, and relative efficiency. Traditional linear tests are targeted broadly so as to have a suitable number of items with target information at every proficiency level. Item-level CATs of the type used in this study select highly informative items based on the examinees' constantly updated proficiency estimates. Differences found between linear and CAT formats demonstrate the impact of each assessment modality on examinee outcomes. To minimize confusion given the multiple levels and conditions in this study, the following shorthand notation will be used to indicate various conditions in this paper: *Linear-Real* denotes the linear test format using real data; *Linear-Sim* denotes the linear test format using simulated data; *CAT-Real* denotes the computer-adaptive testing format using real data; *CAT-Sim* denotes the computer-adaptive testing format using simulated data.

4.2.1 Proficiencies: Real Data Compared to Simulated Data

As one would expect, within both CAT and linear formats, real and simulated datasets differed in their proficiency estimation results. Tables 4.9 – 4.12 summarize proficiency values for the full matrix simulated data as well as complete comparison tables for the three test length conditions. Figures 4.4 – 4.6 demonstrate these results visually. Table 4.9 refers to the full, simulated matrix proficiency overall summaries, which demonstrate that the mean of 0.000 and SD of 1.001 are similar to the real full dataset values of -0.003 and 1.014, respectively (see Table 4.7). Overall standard error (*SE*) values are smaller for the simulated data, with a mean of 0.034 versus the full synthetic matrix SE mean of 0.062. The largest group *SE* differences between the original synthetic matrix and the simulated data were found in the lowest and highest groups, Group 1 and Group 10. Differences in standard deviation values were also largest in Group 1 and Group 10.

25-Item Condition. Table 4.10 contains a summary of the 25-item proficiency estimates. Figure 4.4 graphically illustrates the results of the 25-item test length. In the graph, proficiency means across groups can be seen to follow each other closely across groups within each test condition. Some between-condition deviation can be noted in the higher ability groups, such as Groups 9 and 10 where real and simulated data showed increasingly divergent values. Differences between real and simulated data for the linear

Table 4.9

Full Simulated Matrix Proficiency Estimate $\left(\hat{\theta}\right)$ Statistics (n=500,000)

Group	$\bar{\hat{\theta}}$	$SD\left(\hat{\theta}\right)$	$SE\left(\hat{\theta}\right)$
1	-1.913	0.663	0.066
2	-0.887	0.226	0.004
3	-0.720	0.093	0.006
4	-0.246	0.000	0.001
5	-0.163	0.175	0.022
6	0.246	0.000	0.001
7	0.283	0.121	0.017
8	0.740	0.000	0.006
9	0.985	0.230	0.066
10	1.679	0.259	0.153
ALL	0.000	1.001	0.034

Figure 4.4. Proficiency by Group for 25-Item Test

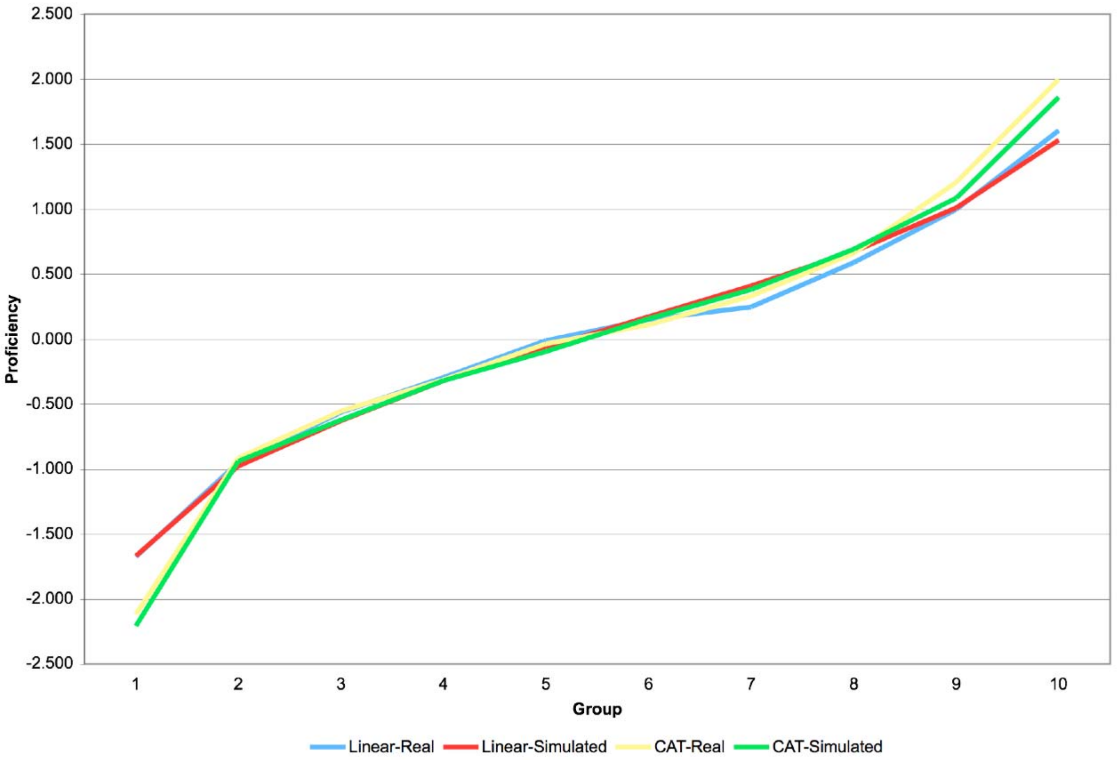


Table 4.10

Examinee Proficiency Estimates ($\hat{\theta}$) for 25-Item Tests

	Linear				CAT			
	Real		Simulated		Real		Simulated	
Group	$\bar{\hat{\theta}}$	$SE\left(\hat{\theta}\right)$	$\bar{\hat{\theta}}$	$SE\left(\hat{\theta}\right)$	$\bar{\hat{\theta}}$	$SE\left(\hat{\theta}\right)$	$\bar{\hat{\theta}}$	$SE\left(\hat{\theta}\right)$
1	-1.673	0.461	-1.668	0.445	-2.118	0.410	-2.208	0.350
2	-0.948	0.339	-0.976	0.342	-0.915	0.167	-0.939	0.164
3	-0.563	0.310	-0.627	0.315	-0.554	0.149	-0.622	0.151
4	-0.294	0.294	-0.319	0.292	-0.313	0.144	-0.319	0.145
5	-0.010	0.307	-0.068	0.301	-0.037	0.151	-0.095	0.148
6	0.152	0.318	0.172	0.321	0.111	0.160	0.156	0.166
7	0.247	0.327	0.410	0.345	0.331	0.181	0.383	0.186
8	0.590	0.369	0.674	0.380	0.662	0.220	0.694	0.225
9	1.000	0.426	1.010	0.429	1.207	0.317	1.086	0.295
10	1.605	0.533	1.531	0.519	1.997	0.471	1.860	0.441
ALL	0.011	0.368	0.014	0.369	0.037	0.237	0.017	0.227

test condition are trivial, demonstrating notably close fitting lines. Referring to Table 4.10, within the 25-item test format the overall linear real and simulated proficiency estimates were quite similar at 0.011 and 0.014, respectively. Standard error (*SE*) values for both types of linear data were nearly identical overall at 0.368 for real data and 0.369 for simulated data. The CAT values for real and simulated data, however, differed by 0.020, with the simulated data having the smaller mean of 0.017 compared to 0.037 for the real data mean. Within the groups, for linear data, Group 7 demonstrated the largest mean proficiency difference between real (0.247) and simulated (0.410) conditions at 0.163. The smallest linear condition real-to-simulated mean proficiency difference was found for Group 1 at 0.005, with -1.673 for real data and -1.668 for simulated data. For the CAT format condition, the largest mean proficiency difference on the 25-item test was found for Group 10 at 0.137, with real data at 1.997 and simulated data at 1.860.

The smallest CAT condition mean difference was found for Group 4 at 0.006. Overall mean *SE* values within the CAT condition were comparable except for Groups 1, 9, and 10, which differed by 0.060, 0.022, and 0.030, respectively. The real data had a slightly higher overall mean *SE* of 0.237 versus the simulated data mean *SE* of 0.227.

35-Item Condition. For the 35-item test, Figure 4.5 shows that the real and simulated data are not identical in that the linear condition lines are not completely overlapping. In particular, linear Groups 2, 8, and 9 show the largest distances between lines. Line distances between real and simulated data for CAT tests are smaller than for linear, but do not overlap completely, with Group 8 showing a clear gap between lines. Linear-Real to Linear-Sim results are somewhat erratic with very large and very small differences found (see Table 4.11). The overall proficiency mean for the real data linear test was -0.034, whereas the simulated data mean was 0.012, resulting in an absolute difference of 0.046. This linear test difference is much higher than the CAT real to simulated difference of 0.005 (0.016 – 0.011). As expected, the calculated *SE* of proficiencies across datasets decreased for the middle groups and increased at the extremes. The calculated *SE* for that group differed by 0.010.

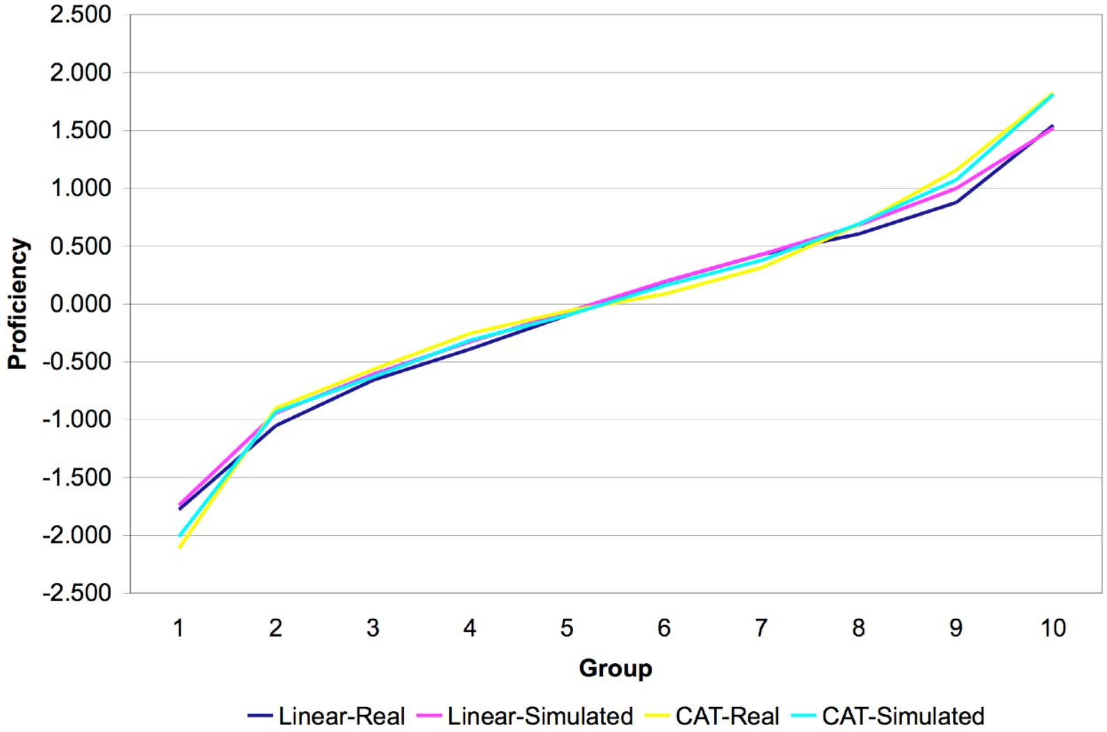
For the linear format, the simulated data produced lower standard errors across the groups than did the real data. For the CAT format condition, the real data *SE*s were very slightly smaller than the simulated data *SE*s, except in the higher and lower groups where the real data produced larger *SE* results than did the simulated data. Group mean proficiency differences for the 35-item test were rather large, with Group 9 in the linear format condition showing the largest difference between real and simulated data at 0.123. Within the linear format, Group 2 also showed a large 0.109 mean proficiency difference, while Group 7 showed a notably small difference of 0.001. The real versus simulated CAT condition showed no real pattern agreement with the linear condition on the mean differences within groups with the exception of Group 9 which had the second largest difference at 0.083, similar to the linear condition. The largest CAT condition group difference was found for Group 1 at 0.101, with a CAT-Real condition value of -2.113 and a CAT-Sim condition value of -2.012. The smallest group difference of 0.005 for the 35-item CAT condition was found for Group 8, with 0.685 for CAT-Real and 0.690 for CAT-Sim.

Table 4.11

Examinee Proficiency Estimates ($\hat{\theta}$) for 35-Item Tests

	Linear				CAT			
	Real		Simulated		Real		Simulated	
Group	$\bar{\hat{\theta}}$	$SE\left(\hat{\theta}\right)$	$\bar{\hat{\theta}}$	$SE\left(\hat{\theta}\right)$	$\bar{\hat{\theta}}$	$SE\left(\hat{\theta}\right)$	$\bar{\hat{\theta}}$	$SE\left(\hat{\theta}\right)$
1	-1.780	0.391	-1.741	0.380	-2.113	0.379	-2.012	0.284
2	-1.054	0.302	-0.945	0.291	-0.904	0.143	-0.934	0.143
3	-0.659	0.281	-0.608	0.276	-0.569	0.130	-0.632	0.133
4	-0.390	0.273	-0.328	0.259	-0.256	0.127	-0.315	0.129
5	-0.100	0.279	-0.071	0.285	-0.063	0.133	-0.100	0.132
6	0.190	0.309	0.193	0.311	0.084	0.143	0.158	0.149
7	0.426	0.340	0.427	0.340	0.313	0.161	0.376	0.165
8	0.605	0.370	0.680	0.382	0.685	0.198	0.690	0.199
9	0.875	0.411	0.998	0.430	1.155	0.268	1.072	0.255
10	1.543	0.510	1.519	0.506	1.822	0.374	1.809	0.369
ALL	-0.034	0.347	0.012	0.346	0.016	0.206	0.011	0.196

Figure 4.5. Proficiency by Group for 35-Item Test



50-Item Condition. Table 4.12 shows that for the 50-item linear condition, the overall mean proficiency differed by 0.005 between real and simulated data types. Within the CAT condition, the absolute difference between real and simulated data was 0.008. The Figure 4.6 graphically illustrates the differences between data types for each testing condition. While the linear test condition performed similarly for real and simulated data in the lower proficiency groups, the means diverged at Groups 3 and 10. Within the CAT condition, the line distance is small in the lower proficiency groups, but the distance increases from Groups 6 to 8, but converges again in the higher proficiency groups. For groups within the linear format, Group 3 demonstrated the largest difference between real and simulated data at 0.135, with Linear-Real being 0.486 and Linear-Sim being 0.621. In the linear format condition, the smallest group difference was found for Group 5 at 0.020. Within the CAT format condition, the group with the largest mean difference was Group 6 at 0.094, and the smallest difference was found for Group 10 at 0.002. The overall mean *SE* values for the real and simulated data within the linear format condition were equal at 0.289, while the CAT condition *SE* values were also very similar at 0.172 for the real data and 0.170 for the simulated data. The largest *SE* values across the proficiency groups were found for the more extreme proficiency groups, (i.e., Groups 1 and 10).

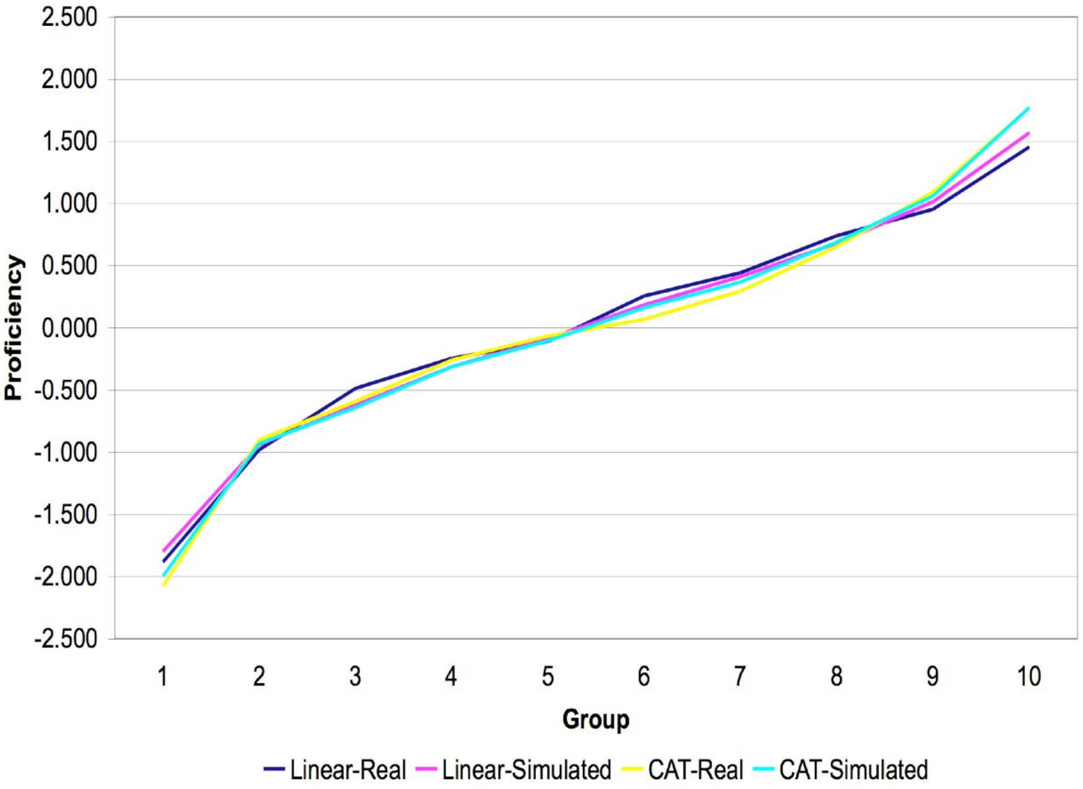
Table 4.12

Examinee Proficiency Estimates ($\hat{\theta}$) for 50-Item Tests

	Linear				CAT			
	Real		Simulated		Real		Simulated	
Group	$\bar{\hat{\theta}}$	$SE\left(\hat{\theta}\right)$	$\bar{\hat{\theta}}$	$SE\left(\hat{\theta}\right)$	$\bar{\hat{\theta}}$	$SE\left(\hat{\theta}\right)$	$\bar{\hat{\theta}}$	$SE\left(\hat{\theta}\right)$
1	-1.883	0.348	-1.797	0.323	-2.075	0.265	-1.997	0.237
2	-0.979	0.239	-0.938	0.235	-0.901	0.125	-0.929	0.125
3	-0.486	0.226	-0.621	0.224	-0.590	0.115	-0.641	0.117
4	-0.244	0.190	-0.313	0.203	-0.260	0.114	-0.311	0.115
5	-0.108	0.235	-0.088	0.238	-0.065	0.120	-0.105	0.119
6	0.257	0.256	0.185	0.255	0.068	0.127	0.162	0.133
7	0.444	0.287	0.411	0.281	0.296	0.142	0.368	0.147
8	0.740	0.330	0.679	0.324	0.653	0.173	0.689	0.177
9	0.953	0.357	1.014	0.366	1.091	0.225	1.059	0.220
10	1.454	0.424	1.569	0.439	1.770	0.314	1.772	0.312
ALL	0.015	0.289	0.010	0.289	-0.001	0.172	0.007	0.170

Real-Simulated Data Proficiency Comparison Summary. Comparing real and simulated data across the three test lengths, proficiency values were found to have slightly smaller values for CAT-Sim data versus its CAT-Real counterpart. The linear data, whether real or simulated, was generally quite comparable across all conditions, with the exception of the Linear-Real 35-item test which was the only negative underestimated overall mean in the study. Overall mean proficiency differences between real and simulated data were the greatest for the 35-item linear test condition at 0.046, by far the largest such difference across conditions. Among the remaining five such linear-real mean proficiency comparisons, the other differences were all 0.020 or smaller, substantially lower values than for the 35-item linear condition. For the standard error values, the more extreme proficiency groups (i.e., 1, 2, 9, and 10), tended to have smaller *SE* values for simulated data than for real data on the 25-item and 35-item tests. Those differences were reduced on the 50-item test. Comparing real to simulated data standard error values, there is more similarity in values for the middle proficiency groups, particularly for the longer test lengths.

Figure 4.6. Proficiency by Group for 50-Item Test



4.2.2 Proficiencies: Linear and CAT Format Comparisons

As in the previous section, a full comparison of the linear and CAT modalities on real and simulated data for proficiency estimation is summarized in Table 4.10 for the 25-item tests, Table 4.11 for the 35-item tests, and Table 4.12 for the 50-item tests. Likewise, Figures 4.4, 4.5, and 4.6 graphically illustrate the proficiency mean values by group.

25-Item Condition. Referring to Figure 4.4, visually comparing values for real data within linear and CAT formats showed that the results were not equal. Real data in Groups 2 through 7 are similar on both linear and CAT tests, but the differences between the two testing formats increase for Groups 8 through 10. Group 1 also shows a large distance between the real data plots. For simulated data, the distance between lines increases notably for both Group 1 and Group 10, the extreme proficiency values. The simulated data lines are otherwise fairly close together across the remaining groups. As can be observed in Table 4.10, for the 25-item test the largest group difference between linear and CAT proficiencies was found in Group 1, for both real (0.445) and simulated (0.540) data. The second largest linear-to-CAT condition difference was found for Group 10 with 0.392 for real data and 0.329 for simulated data, much larger differences than were found for the real-to-simulated data comparisons noted previously. The smallest linear-CAT group difference of 0.009 in mean proficiency estimates for the 25-item linear test condition was found for Group 3.

For the simulated data, the proficiency estimates were found to have a linear versus CAT difference of 0.000 for Group 4. The overall mean proficiency difference between Linear-Real and CAT-Real conditions is 0.026, and the Linear-Sim to CAT-Sim difference is 0.003. Groups with the largest linear-to-CAT proficiency estimate differences were found to have larger *SE* differences as well. Comparing overall means between linear and CAT formats finds rather large *SE* differences: For Linear-Real compared to CAT-Real the overall mean difference was 0.131; and for Linear-Sim compared to CAT-Sim, the overall difference was 0.142. For the Linear-Real condition, the smallest *SE* difference was found in Group 1 at 0.051; for the Linear-Sim condition, the smallest difference was found in Group 10 at 0.078. The group with the largest linear-to-CAT difference was Group 2, for both real (0.172) and simulated (0.178) data. In summary, for both real and simulated data, the group with the largest linear-to-CAT proficiency difference was Group 1, with Group 10 having the second-largest difference. For both real and simulated data, the group with the largest *SE* difference was Group 2. For real data, Group 1 had the smallest *SE* difference, and for simulated data, Group 10 had the smallest *SE* difference.

35-Item Condition. In the 35-item test length condition, Figure 4.5 shows that the linear and CAT formats are similar in the middle groups but distances between the lines are greatest in Groups 1 and 10, the lowest and highest proficiency groups. On the 35-item tests, the linear-CAT difference was largest for Group 1 in the real data condition, and largest for Group 10 in the simulated data condition. Again, this outcome is in keeping with expectations because of the more extreme proficiency values in the distribution tails. Similarly, reviewing Table 4.11, linear versus CAT proficiency differences for real data were highest for Group 1 (0.333) with large differences also found for Group 9 (0.280) and Group 10 (0.279). The smallest linear-to-CAT difference found for real data was for Group 5 at 0.037. Similar to its real data counterpart, simulated data linear-to-CAT differences were highest for Group 10 (0.290) with the next highest value in Group 1 (0.271). The smallest difference between simulated data proficiency values was found for Group 8 at 0.010. Overall mean linear-CAT proficiency differences were 0.050 for real data and 0.001 for simulated data. The smallest linear-to-CAT real data *SE* difference was found for Group 1 at 0.012, and the largest difference was found in Group 7 at 0.179. The smallest difference between simulated data *SE* values was found for Group 1 at 0.096, whereas the largest simulated data *SE* difference was found for Group 8 at 0.183. The overall linear-to-CAT proficiency difference for real data was 0.050, and for simulated data, the difference was 0.001. The overall linear-to-CAT *SE* difference was 0.141 for real data and 0.150 for simulated data. In summary, the largest mean differences for linear versus CAT format tests were found for the extreme proficiency Groups 1 and 10. Comparing within condition, the standard error associated with these more extreme proficiency values was also highest for these groups. Overall mean linear-CAT difference calculations show that real data had larger *SE* values than did simulated data.

50-Item Condition. Referring again to Figure 4.6 for the 50-item condition graph, the real data condition showed the largest distances between linear and CAT lines. In Group 1, the distances between CAT and linear conditions are minimal, but increase to a larger degree in Group 10 for both real and simulated datasets. For the simulated data, linear and CAT lines overlap well for all Groups except 10 and, to a lesser degree, Group 1. In Table 4.12, the largest difference between Linear-Real and CAT-Real values was found for Group 10 at 0.316. Likewise, for Linear-Sim to CAT-Sim differences, Group 10 again was found to have the highest value at 0.203. Overall, however, differences between the two test format conditions, linear and CAT, were relatively small in comparison to the two shorter test length conditions (i.e., 25-item and 35-item tests). The smallest linear-CAT group difference was found for Group 4, at 0.016 for real data and 0.002 for simulated data. The overall mean proficiency difference between the two test conditions was 0.016 for real data and 0.003 for simulated data. Group 8 was found to have the largest *SE* difference between linear and CAT test formats, at 0.157 for real data and 0.147 for simulated data. For real data, the smallest *SE* difference was found in Group 4 at 0.076. For simulated data, the smallest *SE* difference was found for Group 1 at 0.086, although Group 4 was quite similar at 0.088. Overall, the mean SE differences between the linear and CAT formats was 0.117 for real data and 0.119 for simulated data. In summary, for the 50-item test condition, the proficiency extreme Group 10 provided the largest linear-to-CAT difference between means, and Group 4 provided the smallest linear-CAT differences for both data types. Group 8 was found to have the largest *SE* difference for both data conditions, and Group 4 had the smallest *SE* difference for real data and the second smallest SE difference for simulated data. Group 1 had the smallest *SE* difference for simulated data. Overall, the CAT format provided much smaller *SE* values than did the linear format.

Linear-CAT Proficiency Comparison Summary. Across all test length conditions, the linear-to-CAT comparisons of proficiency values found that Groups 1 and 10 had the largest differences. Within test-length condition, the smallest proficiency differences varied by group, with no clear pattern emerging. Overall mean standard error values were higher for linear condition than for CAT condition. The smaller *SE* values for CAT were results one would expect, given that CATs are able to select maximally informative items based on constantly updated proficiency estimates. The selection of a maximally informative item is unlike a linear test which administers the same items to every examinee regardless of proficiency. Given that item information is related to the standard error, one particular strong point of the CAT format is that it minimizes standard error by maximizing information for each examinee based on the examinee's proficiency. Linear tests are unable to utilize this valuable technique. In this study, linear format test items were chosen randomly from stratified bins. Consequently, the 35-item tests behaved in unpredictable ways, perhaps because the items chosen were not sufficiently informative for all examinees as discussed in section 4.1.

4.2.3 Bias and RMSE: Real Data Compared to Simulated Data

Bias and RMSE statistics were used to illustrate differences in proficiency estimation from the true estimates. Bias is a directional measure, while RMSE is non-directional. For example, a positive bias is indicative of estimated proficiencies that are larger than the true proficiencies, an outcome that is often termed overestimation. RMSE accounts for both bias and the associated variance (or precision) and, therefore, a given RMSE value will not necessarily mirror its bias value. Bias and RMSE values for each condition by proficiency group are included in Table 4.13 (25 items), Table 4.14 (35 items), and Table 4.15 (50 items).

25-Item Condition. Table 4.13 summarizes the 25-item condition bias and RMSE values. For both real and simulated data conditions of the 25-item tests, Groups 2 and 4 consistently underestimated proficiency, as noted by the negative bias values. For the Linear-Real condition, the smallest absolute bias value of 0.004 was found for Group 9, and the largest absolute bias of 0.226 was found for Group 6. Within the 25-Item

Table 4.13

Bias and RMSE for 25-Item Test

	LINEAR				CAT			
	Real		Simulated		Real		Simulated	
Group	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
1	0.206	0.599	0.157	0.612	-0.205	0.559	-0.099	0.423
2	-0.166	0.266	-0.054	0.351	-0.047	0.235	-0.051	0.178
3	0.158	0.378	-0.017	0.295	0.151	0.231	0.045	0.159
4	-0.036	0.354	-0.007	0.296	-0.102	0.266	-0.035	0.143
5	-0.130	0.253	-0.009	0.295	-0.026	0.239	0.025	0.174
6	0.226	0.397	0.007	0.310	0.067	0.246	-0.036	0.142
7	-0.126	0.343	0.016	0.341	0.039	0.215	0.042	0.203
8	0.029	0.279	0.011	0.389	-0.028	0.170	-0.013	0.206
9	0.004	0.314	-0.009	0.446	0.208	0.302	0.050	0.312
10	-0.031	0.243	0.072	0.433	0.341	0.463	0.275	0.493
ALL	0.013	0.357	0.017	0.389	0.040	0.315	0.020	0.270

Linear-Sim condition, the smallest bias absolute value was tied for Group 4 (-0.007) and Group 6 (0.007). The largest bias value within the Linear-Sim condition was found for Group 1 at 0.157. Within the Linear-Real condition, the largest RMSE value was for Group 1 at 0.599, and Group 1 also had the second-largest bias value. This result signifies that Group 1 estimates were consistently too large but they also varied notably within that overestimation. For the Linear-Sim data, the largest RMSE was 0.612, also for Group 1. The smallest RMSE values were found for Groups 3 and 5, both 0.295, and Group 4 was nearly identical at 0.296. Because Group 4 had the smallest absolute bias value (along with Group 6) and also had a low RMSE, one can conclude that both the proficiency variance and estimation bias were small values. The largest negative bias values were found for Group 2, for both Linear-Real (-0.166) and Linear-Sim (-0.054) conditions. Therefore, for Group 2, in both linear conditions, the proficiency estimates were lower than the true proficiency. For the linear condition, however, Group 2 estimates varied less in the real data condition than in the simulated data condition as noted by the smaller real data RMSE (0.266 versus 0.351). The largest linear format RMSE was found in the simulated data condition in Group 1. Though the Group 1 Linear-Sim bias was smaller than in the real data condition, the RMSE was larger, indicating greater variance within Group 1. For the linear test format, the overall mean bias was larger for the simulated data (0.017) than for the real data (0.013), and the overall RMSE was also larger for the simulated data (0.389) than for the real data (0.357).

The 25-Item CAT format showed greater bias differences between real and simulated data, compared the linear format, with overall mean bias differing by 0.020 (0.040 and 0.020, respectively). The largest CAT-Real condition absolute bias was found for Group 10 (0.341). Likewise, the CAT-Sim condition was also found to have its highest absolute value for bias in Group 10 (0.275). The CAT-Real condition had as its smallest bias absolute value -0.026 for Group 5, whereas the smallest value for the CAT-Sim condition was found in Group 8 at -0.013. Mean CAT RMSE values were 0.315 for real data and 0.270 for simulated data, resulting in a mean difference of 0.045, combined with the smaller overall bias, this result indicates that the simulated data better represent the true proficiency values. The largest RMSE value for the CAT-Real condition was 0.559, which was found in Group 1. The smallest CAT-Real RMSE was found for Group 8 (0.170), which also had the smallest absolute bias value of -0.028. In the CAT-Sim condition, the largest RMSE value was found for Group 10 (0.493), while the smallest RMSE value was found for Group 6 (0.142) with Group 4 at nearly the same value (0.143). Given that Group 10 had the largest bias value, and Groups 4 and 6 had the two smallest values, this result implies that the variation across CAT-Sim values was consistent from group to group. By comparison, Groups 2 and 9 had nearly identical bias values (-0.051 and 0.050, respectively) but vastly different RMSE values (0.178 and 0.312, respectively). This result indicates that Group 9 had much more score variation than did Group 2.

Table 4.14

Bias and RMSE for 35-Item Test

	LINEAR				CAT			
	Real		Simulated		Real		Simulated	
Group	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
1	0.032	0.558	0.141	0.522	-0.182	0.493	-0.078	0.338
2	-0.074	0.314	-0.054	0.292	-0.047	0.203	-0.048	0.152
3	-0.108	0.314	-0.007	0.259	0.153	0.216	0.054	0.137
4	-0.178	0.357	-0.004	0.268	-0.061	0.211	-0.048	0.119
5	-0.072	0.324	-0.011	0.279	-0.002	0.248	0.032	0.157
6	0.191	0.386	0.009	0.298	-0.008	0.203	-0.044	0.122
7	-0.117	0.440	0.017	0.339	0.009	0.181	0.049	0.183
8	0.004	0.276	0.015	0.391	0.004	0.160	-0.022	0.175
9	0.097	0.486	-0.014	0.445	0.111	0.224	0.048	0.278
10	-0.090	0.369	0.061	0.430	0.206	0.374	0.198	0.398
ALL	-0.032	0.391	0.015	0.363	0.018	0.269	0.014	0.226

35-Item Condition. For the 35-item tests (see Table 4.14) the overall mean bias for the Linear-Sim condition was 0.015, whereas the overall mean bias for the Linear-Real condition was -0.032, reversed in sign and notably larger than the Linear-Sim condition. The largest group absolute bias was found for Group 6 at 0.191. Within the same condition, the smallest absolute bias was found for Group 8 at 0.004. For the Linear-Sim condition, the largest absolute bias was found for Group 1 at 0.141, while the smallest absolute bias value was found for Group 4 at -0.004, which is a notably smaller value than was found for the other Linear-Sim groups in this condition. Within the 35-

item Linear format condition, Group 1 was found to have the largest RMSE value for both the real and simulated data conditions. The smallest Linear-Real RMSE value of 0.276 was found for Group 8, which also had the smallest bias value (0.004). For the Linear-Sim condition, the smallest RMSE of 0.259 was found in Group 3, which also had the second-smallest bias value in that condition. The Linear-Real condition had larger overall bias and RMSE values (-0.032 and 0.391, respectively) than did the Linear-Sim condition (0.015 and 0.363, respectively).

For the 35-item CAT testing format, the real data condition had the largest absolute bias in Group 10 at 0.206. In the same CAT-Real condition, Group 5 showed the lowest absolute bias at -0.002. The smallest positive bias value was found in Group 8 at 0.004; correspondingly, the smallest RMSE value for this condition was found in Group 8 at 0.160. The largest CAT-Real negative bias was found for Group 1, and the largest positive bias was found for Group 10. The largest CAT-Real RMSE value was found for Group 1 at 0.493. The 35-item CAT- Sim format, like the Real data format, was found to have the largest absolute bias value of 0.198 for Group 10. Group 8 showed the smallest absolute bias at -0.022. The largest CAT-Sim RMSE value was found for Group 10 (0.398) and the smallest RMSE was found for Group 4 (0.119). Overall, for the 35-item test, the observed RMSE general pattern is similar to the pattern found on the 25-item test. RMSE values were highest at the group extremes and lowest in the middle of the proficiency groups. In addition, RMSE values were smaller for simulated data than they were for real data. The overall mean Linear-Real RMSE of 0.391 is notably higher than the overall Linear-Sim RMSE of 0.363. Similarly, the overall CAT-Real RMSE of 0.269 is larger than its CAT-Sim counterpart, with an overall mean of 0.226. The RMSE statistic for the CAT-Real groups reached a maximum value of 0.493 for Group 1, whereas the highest RMSE group value for the CAT-Sim condition occurred in Group 10. Among the bias and RMSE statistics by group for the 35-item test, 7 out of 8 conditions attained their maximum RMSE or bias values in group 1 or group 10. Only the Linear-Real data condition did not have its maximum bias in either Group 1 or Group 10.

Table 4.15

Bias and RMSE for 50-Item Test

	LINEAR				CAT			
	Real		Simulated		Real		Simulated	
Group	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
1	-0.011	0.448	0.115	0.426	-0.138	0.386	-0.061	0.281
2	-0.057	0.286	-0.047	0.235	-0.041	0.187	-0.044	0.131
3	0.133	0.304	-0.001	0.213	0.133	0.199	0.055	0.121
4	0.059	0.248	0.002	0.219	-0.048	0.165	-0.054	0.102
5	-0.016	0.230	-0.015	0.233	-0.049	0.210	0.036	0.143
6	0.123	0.308	0.011	0.244	0.033	0.202	-0.049	0.106
7	0.062	0.295	0.014	0.285	-0.019	0.143	0.058	0.165
8	-0.194	0.422	0.013	0.326	-0.056	0.145	-0.033	0.147
9	0.087	0.362	-0.010	0.390	0.006	0.259	0.045	0.246
10	-0.011	0.317	0.050	0.382	0.195	0.360	0.143	0.325
ALL	0.018	0.329	0.013	0.305	0.002	0.239	0.010	0.192

50-Item Condition. For the 50-item Linear-Real condition, the largest bias absolute value was found for Group 8 at -0.194, and the smallest bias absolute value was found for Group 1 and Group 10, both -0.011 (see Table 4.15). Group 5 was found to have the second smallest absolute bias of -0.016. The largest Linear-Real RMSE value was found for Group 1 at 0.448, and the smallest RMSE was found for Group 5 at 0.230. The large RMSE value for Group 1 along with the smallest bias values indicates that there was a large amount of variation in proficiencies. In contrast, within the Linear-Sim condition, the group with the largest absolute bias was Group 1 at 0.115. Linear-Sim format Group 3 had the smallest absolute bias at -0.001, which is the smallest group bias value across all conditions and test lengths. In the Linear-Sim condition, the largest RMSE value was found for Group 1 at 0.426 and the smallest RMSE was found for Group 3 at 0.213. For the Linear-Sim condition, the bias and RMSE values were matched to each other for highest and lowest values in their condition, indicating relatively consistent variation across Linear-Sim groups. For the 50-item linear test format, the Linear-Real overall mean bias was 0.018; the Linear-Sim condition overall mean bias was somewhat smaller at 0.013. In the linear format, the overall mean RMSE for the Real data condition was 0.329, and the overall mean RMSE for the Sim data condition was a smaller 0.305, resulting in a difference of 0.024. Therefore, in the linear format the real data had greater overall bias and RMSE values than its simulated data counterparts.

For the CAT-Real groups, the largest bias absolute value was found for Group 10 at 0.195, whereas the smallest absolute bias was found for Group 9 at 0.006. The largest CAT-Real RMSE was found for Group 1 at 0.386; the smallest RMSE value was found for Group 7 at 0.143.   The CAT-Sim condition showed the largest absolute bias in Group 10 (0.143), the same group as in the CAT-Real condition. The smallest CAT-Sim absolute bias value was found in Group 8 at -0.033. The overall mean Linear-Sim bias value was 0.013, a smaller value than the Linear-Real mean bias of 0.018. The difference between the overall Linear-Real bias mean and Linear-Sim bias mean was 0.005, which is similar to the 0.004 difference found for the same 25-item test comparison. Unlike the three other conditions of the 50-item test length which had the largest RMSEs for Group 1, the largest CAT-Sim RMSE was found for Group 10 at 0.325. The smallest CAT-Sim RMSE of 0.102 was found for Group 4. In the CAT-Real condition, the overall mean bias was 0.002, whereas the CAT-Sim condition had a slightly higher overall mean bias value of 0.010. The CAT-Real overall mean bias of 0.002 was the lowest overall mean bias of all conditions for all test lengths in this study. As with the other test length conditions, the 50-item test was found to have 7 out of 8 of the maximum RMSE and bias values located in either Group 1 or Group 10. Again, only the Linear-Real bias condition did not follow this pattern, having its maximum bias within Group 8. In addition, for the Linear-Real bias statistic, the minimum values were tied at -0.011 for Groups 1 and 10. Other bias and RMSE minimum values in the 50-item condition occurred primarily in Groups 3 through 7, with a notable exception for the CAT-Real condition where the minimum bias of 0.006 occurred in Group 9.

Summary of Real-Simulated Data Bias and RMSE Comparison. Across conditions, the overall mean bias and RMSE values were generally larger for real data than for simulated data, with the exception of the 25-item linear format, and the 50-item CAT-Real bias which was smaller than its CAT-Sim counterpart. Only the 35-item Linear-Real overall bias had a negative value which was also the second largest overall bias in the study. This result indicates that the 35-item Linear-Real condition consistently underestimated proficiencies. In the groups analysis, Groups 1 and 10 often had the largest bias and RMSE values. The smallest bias and RMSE values typically occurred in Group 3 through Group 8. Simulated data tended to have extreme bias and RMSE values occur together more often than real data. These results indicate less variation in extreme proficiency estimates for simulated data than for real data.

4.2.4 Bias and RMSE: Linear Compared to CAT Formats

25 Item Condition. The overall mean bias value for the 25-item Linear-Real data was 0.013 (Table 4.13). The overall mean bias result for 25-item CAT-Real condition was found to be 0.040, a somewhat larger value than was found for the Linear-Real condition. Similarly, the 25-item Linear-Sim condition overall mean bias was found to be 0.017, while the CAT-Sim data yielded an overall bias of 0.020. Thus, the 25-item CAT condition, regardless of real versus simulated data condition, showed higher mean bias values than did the linear format condition. Generally, this outcome would not be expected due to the strengths of the CAT format. In this case, however, the CAT did a rather poor job of reducing bias in the CAT-Sim high proficiency group, Group 10, which had an absolute bias value of 0.275. For Linear-Real and CAT-Real, there was some bias pattern overlap, with Linear-Real having its smallest absolute bias in Group 8 (0.029), and CAT-Real having its second-smallest absolute bias in the same group (-0.028). The overall mean Linear-Real RMSE was 0.357, whereas the CAT-Real RMSE was somewhat lower at 0.315. Similarly, the Linear-Sim mean RMSE was 0.389 in comparison to the much smaller CAT-Sim mean RMSE of 0.270. Thus, despite the slightly higher bias in the CAT-Sim condition versus the Linear-Sim condition, the score variation within the CAT format was much lower than in the linear format. The largest RMSE value was found in Group 1 for both Linear-Real (0.599) and CAT-Real (0.559) conditions. For the Linear-Sim condition, the largest RMSE was also found in Group 1 (0.612), but for the CAT-Sim condition, the largest RMSE was found in Group 10 (0.493). The second-largest RMSE, however, was found for Group 1 in the CAT-Sim condition. The 25-item CAT format condition demonstrated notably smaller overall

RMSE values than did its linear test counterparts, indicating improved recovery of the original proficiency estimates for the CAT format.

35 Item. For the 35-item condition, the CAT-Real condition resulted in an overall mean bias of 0.018 (see Table 4.14). In contrast, the Linear-Real condition resulted in a negative overall bias value of -0.032. Comparing bias values between linear-to-CAT groups, no correspondences were found among groups for largest or smallest bias values. For the CAT format, the RMSE values were much lower than for their linear test counterparts. For the CAT-Real condition, the overall RMSE mean was 0.269 compared to 0.391 for the Linear-Real condition; for CAT-Sim, the RMSE overall mean was 0.226 compared to 0.363 for Linear-Sim. At 0.391, the overall mean RMSE for the Linear-Real 35-item condition was the largest RMSE value for any data type, format, or test length in this study. This outcome indicates that both bias and variation were large values. The overall Linear-Real mean bias of -0.032 was the second-largest overall bias value in the study. While a smaller value than was found in the Linear-Real condition, the Linear-Sim condition had an overall mean RMSE of 0.363, also a large value. Both Linear-Real and CAT-Real conditions were found to have their smallest RMSE values in Group 8 (0.276 and 0.160, respectively), and their largest RMSE values (0.558 and 0.493, respectively) also occurred in Group 1. This outcome indicates that RMSE results in this condition were based more on data type than test administration format. Because bias and RMSE are related to proficiency estimation, the unusual results from the 35-item tests noted previously continued to be a factor in these results particularly for Linear-Real data. Overall, both bias and RMSE values were larger for the linear tests than those of their CAT counterparts.

50-Item Condition. The 50-item tests resulted in overall bias and RMSE statistics that were the lowest among the three test lengths (Table 4.15). The CAT results in particular resulted in the smallest RMSE values. The Linear-Real overall mean bias was 0.018, while the CAT-Real overall mean bias was the lowest of any in this study at 0.002. The Linear-Sim overall bias mean of 0.013 was similar to the CAT-Sim overall bias mean of 0.010. The largest Linear-Real absolute bias value was for Group 8 at -0.194, whereas the largest CAT-Real absolute bias was found for Group 10 at 0.195. The smallest bias for the Linear-Real data was found for Groups 1 and 10, tied at -0.011, but the smallest absolute bias value of 0.006 for the CAT-Real dataset was found for Group 9. The largest Linear-Sim absolute bias was found in Group 1 at 0.115, and the largest absolute bias found for the CAT-Sim data was found in Group 10 at 0.143. As described previously, the smallest absolute bias for the Linear-Sim groups occurred in Group 3 (-0.001), and the same measure for the CAT-Sim group occurred in Group 8 (-0.033). As has been found in the other two test length conditions, RMSE values within groups are lowest in the middle range of proficiency and highest at more extreme proficiency value groups. The largest overall mean RMSE in the 50-item test format was found for the Linear-Real condition at 0.329, whereas the CAT-Real overall RMSE was a much smaller 0.239, a notable difference of 0.090. The largest RMSE in the Linear-Real condition was found in Group 1 at 0.448, while the largest RMSE found for the CAT-Real data was 0.386, also for Group 1. For the Linear-Real condition, the smallest RMSE of 0.230 was found for Group 5, whereas the smallest CAT-Real RMSE of 0.143 was found in Group 7. Both of these smallest RMSE values were located near the center of the proficiency distribution, but the CAT value was much smaller than the linear value.

For the simulated data, the Linear-Sim condition overall mean RMSE was 0.305 and the overall mean RMSE for the CAT-Sim condition was 0.192, a large difference of 0.113. The overall CAT-Sim RMSE of 0.192 is the lowest overall RMSE value in the study. The largest Linear-Sim RMSE was found for Group 1 at 0.426, whereas for the CAT-Sim condition the largest RMSE of 0.325 was found for Group 10. This between-method difference of 0.101 was one of the largest in the study.

Summary of Linear-CAT Bias and RMSE Comparison. Generally, the CAT condition overall bias and RMSE values were smaller than their linear counterparts in every test length, except for the 25-item test where the CAT condition bias values were larger than the linear bias values. Across all test lengths, the largest RMSE values for real data were found in Group 1. Similarly, Linear-Sim conditions always had their largest bias and RMSE values in Group 1, whereas CAT-Sim conditions always had their largest bias values in Group 10. This outcome implies that simulated data linear and CAT formats were least effective at targeting proficiencies accurately at different parts of the data distribution, the lowest proficiencies for linear and the highest proficiencies for CAT. For real data, the largest RMSE values were found consistently in Group 1 in both linear and CAT test formats, indicating that test format made little difference for real data RMSE, possibly because the proficiency variation was always large at the lowest end of the data distribution.

4.2.5 Item Information: Linear Compared to CAT Formats

As would be predicted from the theory, item information across the study conditions was highest within the CAT format. Typically, an important reason for using a CAT format examination is the ability of adaptive tests to select items with the highest information for any estimated proficiency value. Table 4.16 and Figure 4.7 (25 items), Table 4.17 and Figure 4.9 (35 items), and Table 4.18 and Figure 4.11 (50 items) summarize the results of the mean item information values in both tabular and graphic formats. Figures 4.8, 4.10, and 4.12 show the CAT-linear information differences for the 25, 35, and 50-item tests, respectively.

25-Item Condition. For the 25-item condition, overall mean item information for the Linear-Real condition was found to be 0.295 (Table 4.16). In contrast, the 25-item CAT-Real data yielded much higher mean item information of 0.713. Based on group mean data, the most informative items were found in the middle proficiency distribution, while the most extreme proficiency groups resulted in lower item information. For all four 25-item conditions, the largest information mean was found for Group 4: 0.464 for Linear-Real data and 1.920 for CAT-Real data. Likewise, for this test length, the

Table 4.16

Item Information for 25-Item Test

	Linear		CAT	
Group	Real	Sim	Real	Sim
1	0.188	0.202	0.240	0.328
2	0.347	0.341	1.442	1.496
3	0.417	0.404	1.806	1.743
4	0.464	0.469	1.920	1.898
5	0.426	0.441	1.764	1.824
6	0.395	0.388	1.561	1.453
7	0.373	0.337	1.219	1.162
8	0.293	0.277	0.826	0.793
9	0.220	0.217	0.398	0.460
10	0.141	0.148	0.181	0.205
ALL	0.295	0.294	0.713	0.776

Figure 4.7

Mean Item Information for 25-Item Test

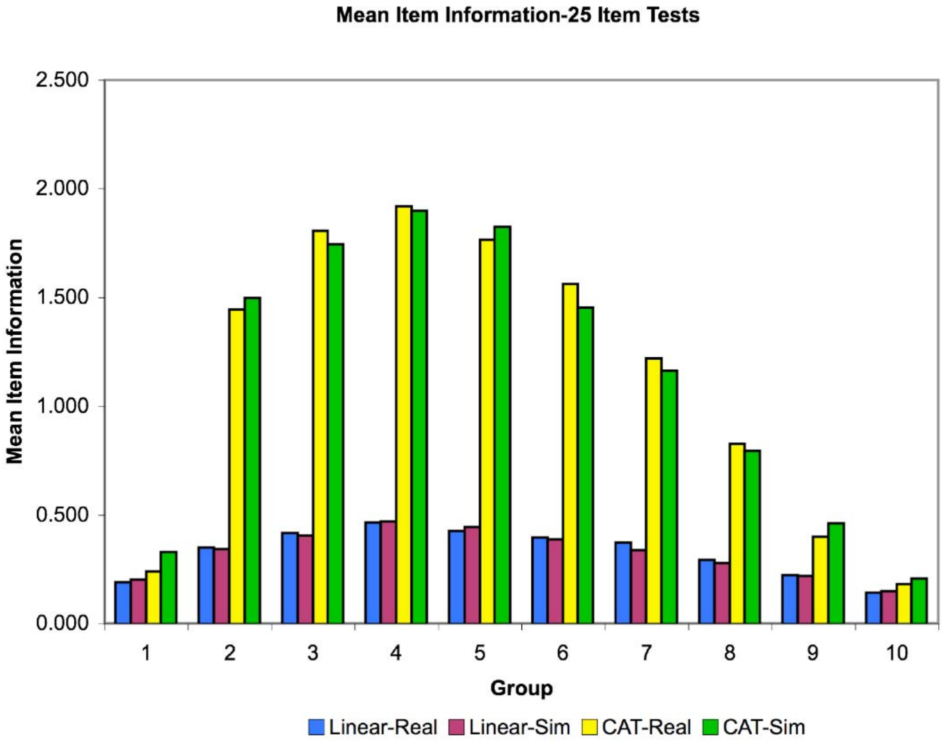
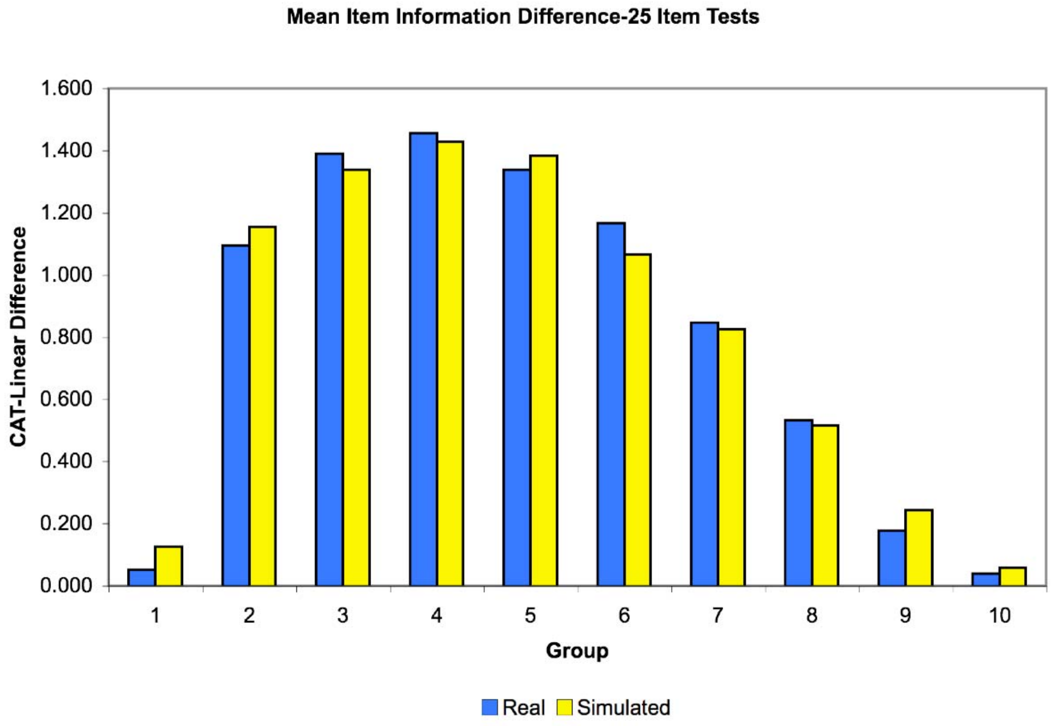


Figure 4.8

CAT-Linear Differences in Item Information – 25 Items



smallest mean information values for all four conditions were found for Group 10: 0.141 for Linear-Real data and 0.181 for CAT-Real data. The CAT values are much higher than linear values for the highest proficiency groups, but this effect is greatly diminished for the lowest proficiency groups where the information means have more similar values regardless of data or administration type. Figure 4.8 graphically illustrates the differences between CAT and linear formats. The real data CAT-to-Linear overall mean difference was 0.418, a smaller overall difference than was found for the simulated data which was 0.482. The larger simulated data difference was due to the CAT-Sim data having the large overall mean information for this test length at 0.776. The largest group Linear-Real condition difference of 1.456 was found for Group 4. Likewise, for simulated data, the largest CAT-to-Linear information difference of 1.429 was also found in Group 4. In contrast, the smallest real data CAT-to-linear difference of 0.040 was found in Group 10, with Linear-Real (0.141) being only somewhat smaller than CAT-Real (0.181). Likewise, for simulated data, the smallest difference of 0.057 was also found in Group 10, again with CAT-Sim (0.205) being larger than Linear-Sim (0.148). For simulated data, the largest CAT-to-linear difference was found for Group 4 at 1.429. Figure 4.8 shows the large differences in item information between linear and CAT format conditions. Both Figure 4.7 and Figure 4.8 illustrate the decline in information at the extreme proficiency group values.

Table 4.17

Item Information for 35-Item Test

	Linear		CAT	
Group	Real	Sim	Real	Sim
1	0.187	0.198	0.201	0.355
2	0.314	0.336	1.388	1.402
3	0.362	0.374	1.696	1.623
4	0.383	0.426	1.782	1.730
5	0.367	0.352	1.613	1.628
6	0.300	0.295	1.389	1.293
7	0.247	0.247	1.098	1.045
8	0.209	0.196	0.726	0.723
9	0.169	0.155	0.397	0.441
10	0.110	0.112	0.204	0.210
ALL	0.238	0.239	0.675	0.746

Figure 4.9

Mean Item Information for 35-Item Test

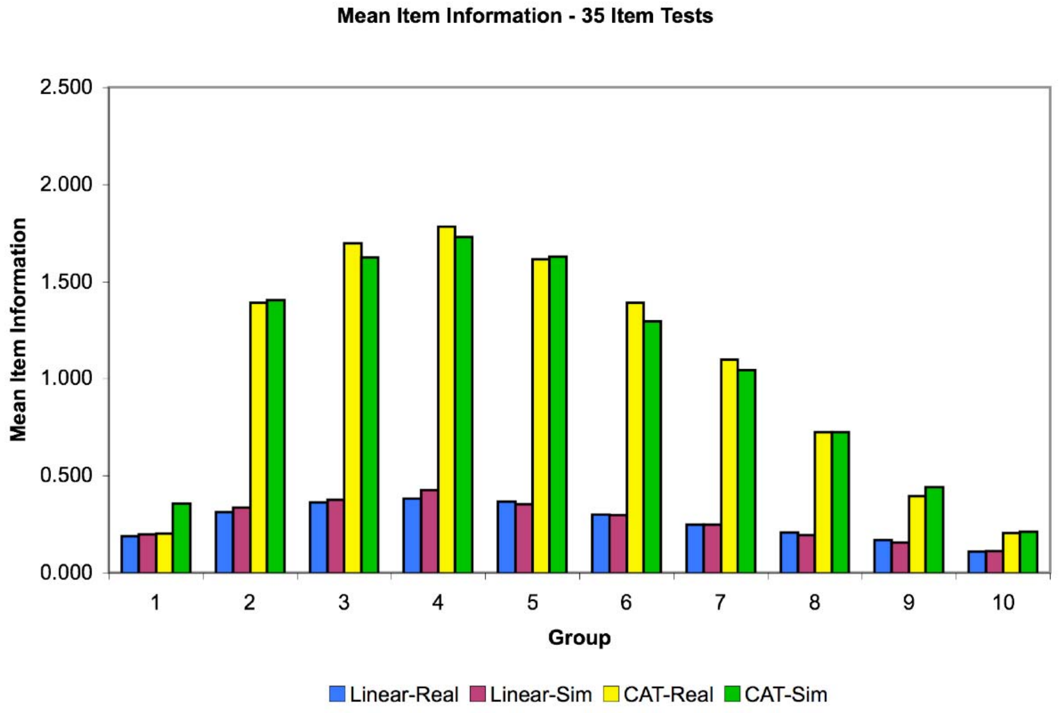
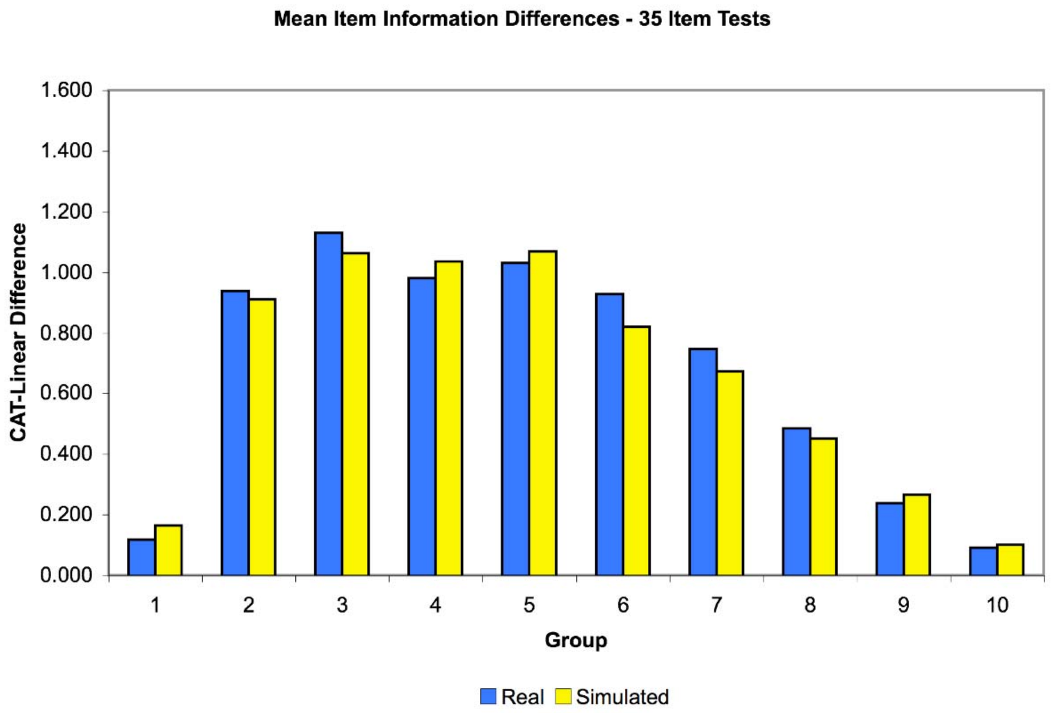


Figure 4.10

CAT-Linear Differences in Item Information – 35 Items



35-Item Condition. For the 35-item tests, the item information results are shown in Table 4.17 and displayed graphically in Figure 4.9. The data indicate that more extreme proficiency groups have lower information, regardless of test format. For the CAT administration format, mean information was greater for all groups, but particularly for Groups 2 through 9 as compared to the minimal differences for the extreme Groups 1 and 2. This outcome is expected because computer-adaptive tests select items based on maximizing information. The Linear-Real overall mean was 0.238, and the CAT-Real overall mean was 0.675. The highest information value for both Linear-Real and CAT-Real conditions was found for Group 4 at 0.383 and 1.782, respectively. Likewise, for the Linear-Sim and CAT-Sim condition, the highest information was found for Group 4, 0.426 and 1.730, respectively. Figure 4.10 illustrates the mean differences between CAT and linear test formats. The largest overall information value for this test length was found for the CAT-Sim condition at 0.746. Therefore, the largest overall mean CAT-Linear difference of 0.507 was found within the simulated data condition. The real data condition overall mean CAT-to-Linear difference was 0.437. The largest CAT-to-Linear group differences were found in Group 4 at 1.399 for real data and 1.304 for simulated data. The smallest CAT-Linear differences for real data were found in Group 1 at 0.014. For the simulated data, the smallest CAT-Linear difference was found in Group 10 at 0.098.

Table 4.18

Item Information for 50-Item Test

	Linear		CAT	
Group	Real	Sim	Real	Sim
1	0.165	0.191	0.284	0.355
2	0.351	0.363	1.288	1.274
3	0.392	0.398	1.521	1.460
4	0.554	0.483	1.533	1.518
5	0.361	0.353	1.392	1.421
6	0.305	0.309	1.233	1.129
7	0.242	0.253	0.989	0.925
8	0.183	0.191	0.668	0.642
9	0.157	0.150	0.396	0.415
10	0.111	0.104	0.203	0.205
ALL	0.239	0.240	0.676	0.691

Figure 4.11

Mean Item Information for 50-Item Test

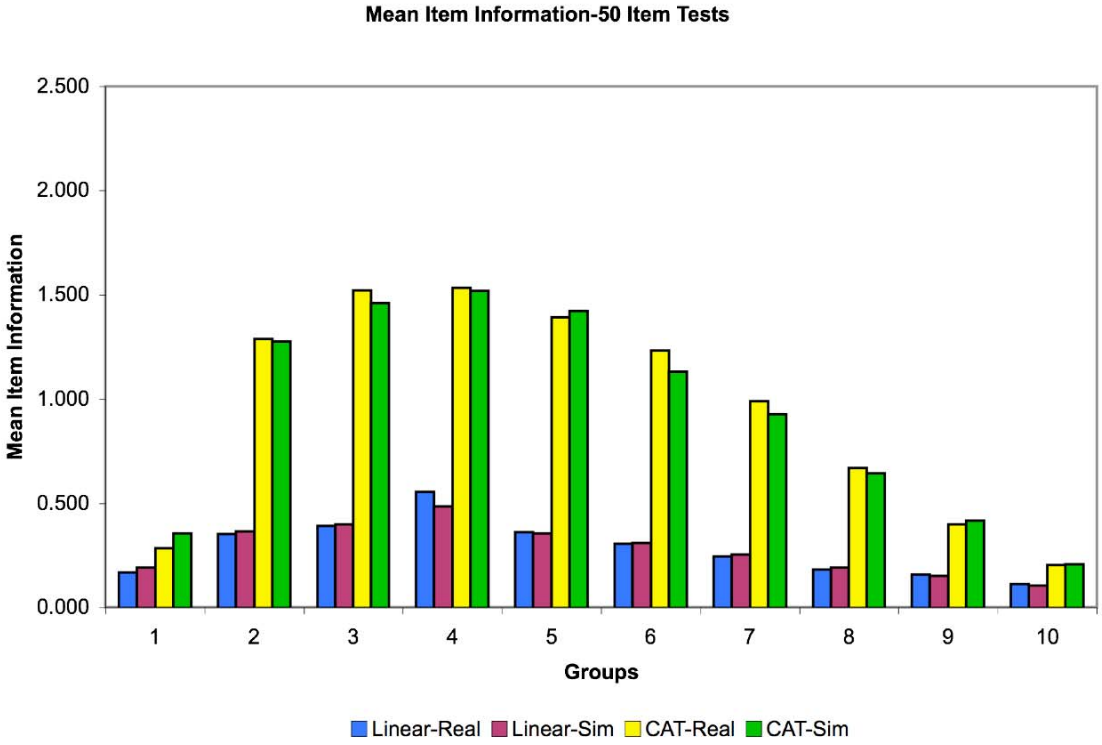
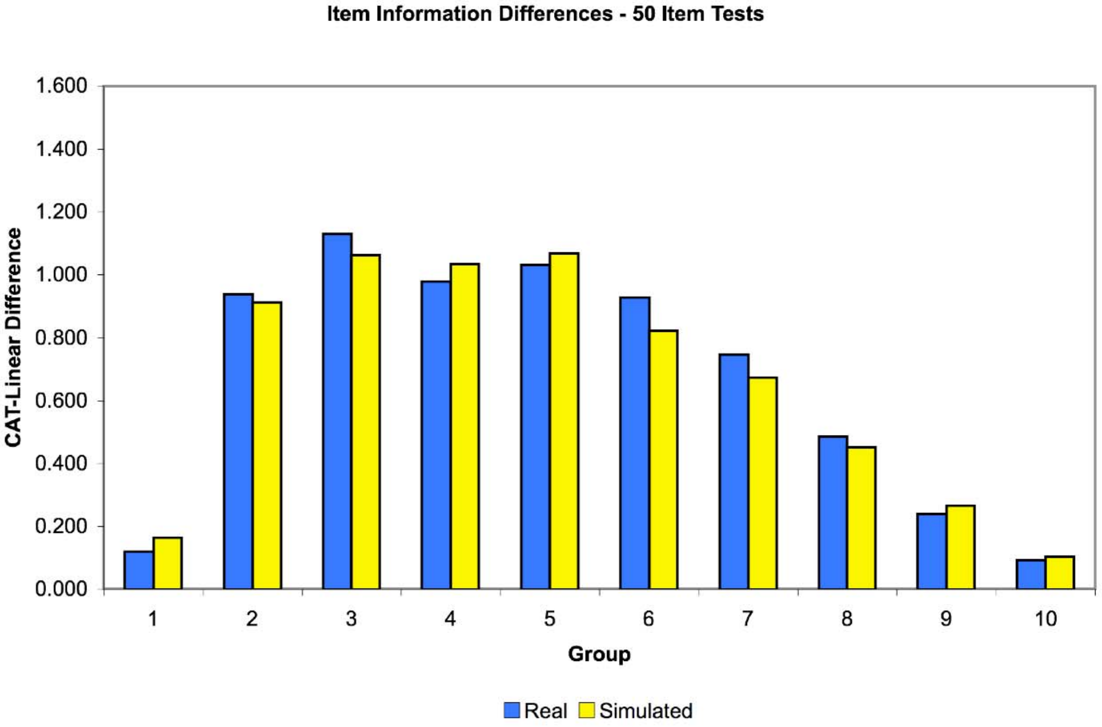


Figure 4.12

CAT-Linear Differences in Item Information – 50 Items



50-Item Condition. The longest test condition of 50 items showed similar patterns to the two shorter tests as shown in Table 4.18 and illustrated in Figure 4.11 and Figure 4.12. Overall, CAT information means were much higher than their linear test counterparts. The Linear-Real overall mean was 0.239, whereas the CAT-Real mean was 0.676. For simulated data, the Linear-Sim overall mean was 0.240, and the CAT-Sim overall mean was 0.691. Once again, within every condition, Group 4 was found to have the highest mean item information. The Group 4 Linear-Real condition information value was 0.554 whereas the CAT-Real condition information was nearly 3 times greater at 1.533. The lowest mean information for the 50-item test was found for Group 10 across all conditions. Following similar patterns to the shorter test lengths, the 50-item test showed the highest mean information for the middle to lower proficiency groups, with the lowest information values being found at the proficiency group extremes.

Linear-CAT Information Summary. Comparing linear and CAT formats on item information, CAT tests showed much larger information values than their linear test counterparts across all study conditions, including data types. Across test lengths, the highest information was found in Group 4 and the lowest information was found in Group 10, with the exception of the 35-item test which had its lowest information in Group 1 for the CAT-Real condition.

4.2.6 Item Information: Real Compared to Simulated Data

Comparing the real to simulated information data, across all conditions, the real and simulated data were found to have similar values, unlike the Linear-to-CAT format comparisons which were highly divergent.

25-Item Condition. For the 25-item condition, the overall mean for the Linear-Real condition was 0.295, and the Linear-Sim condition mean was a nearly identical 0.294 (Table 4.16). The largest information value within the linear groups was found for Group 4 for both real and simulated data. The Group 4 Linear-Real value was 0.464 and the Linear-Sim value was 0.469. Group 4 also constituted the smallest real-to-simulated information difference at 0.005. The largest real-to-simulated data difference was found for Group 7 at 0.036, with the Linear-Real data showing the greater value at 0.373 versus the Linear-Sim value of 0.337. The smallest information values within the linear condition were found for Group 10, for both real (0.141) and simulated (0.148) data. The overall mean absolute difference between Linear-Real and Linear-Sim was calculated to be 0.001. As noted previously, the CAT information values were much larger than their linear test counterparts. The overall CAT-Real mean was 0.713, and the CAT-Sim mean was 0.776. The CAT group with the largest information value was Group 4, for both real (1.920) and simulated (1.898) data. As with the linear format data, Group 4 also showed the smallest difference between CAT-Real and CAT-Sim conditions at 0.022. The group with the largest difference was Group 6 at 0.108 between CAT-Real (1.561) and CAT-Sim conditions (1.453). The overall mean absolute difference between CAT-Real and CAT-Sim was calculated to be 0.063, a much larger difference than the 0.001 difference found for the linear test condition.

35-Item Condition. For the 35-item condition (Table 4.17), the overall information mean for the Linear-Real condition was 0.238, and the overall mean for the Linear-Sim condition was 0.239, which are nearly identical values. Across all conditions within the 35-item tests, Group 4 was found to have the largest information values. The value found for Group 4 in the Linear-Real condition was 0.383, comparable but quite a bit lower than the Linear-Sim value of 0.426. In fact, Group 4 was found to have the largest Linear-Real to Linear-Sim difference at 0.043. The smallest overall information values were found for Group 10 in both real (0.110) and simulated (0.012) data in the linear format condition. The smallest linear format difference between data types was found for Group 7 where there was a 0.000 difference, as both Linear-Real and Linear-Sim had the same value of 0.247. Within the CAT format conditions, the overall CAT-Real mean was 0.675 and the CAT-Sim overall mean was somewhat higher at 0.746. Therefore, the overall difference between CAT-Real and CAT-Sim means was 0.071. For both CAT-Real and CAT-Sim conditions, Group 4 was found to have the largest information values at 1.782 and 1.730, respectively. For the CAT-Real condition, Group

1 had the smallest information value at 0.201, although Group 10 was a similar 0.204. Because of this small Group 1 value, it also had the largest real-to-simulated data difference at 0.154 due to the much larger CAT-Sim value of 0.355. The smallest between-condition difference of 0.003 was found for Group 8, a result of the similar CAT-Real (0.726) and CAT-Sim values (0.723).

50-Item Condition. For the 50-item real-to-simulated data comparison, a comparable pattern was found to the pattern in the 25-item condition (Table 4.18). The overall Linear-Real mean was found to be 0.239 and the Linear-Sim overall mean was a nearly identical 0.240, a difference of only 0.001. The largest group mean was found for Group 4, across every condition: 0.554 for Linear-Real, 0.483 for Linear-Sim, 1.533 for CAT-Real, and 1.518 for CAT-Sim. The largest real-to-simulated data difference in the linear condition was found for Group 4 at 0.071. In the same condition, the smallest group difference was found for Group 6 at 0.004 (0.309-0.305). The smallest linear condition difference was found in Group 10 at 0.111 for Linear-Real and 0.104 for Linear-Sim. Likewise, within the CAT conditions, the smallest information means were found for Group 10 at 0.203 for CAT-Real and 0.205 for CAT-Sim. These small, similar values indicate that Group 10 had the smallest difference between real and simulated data at 0.002. The largest difference between data types was found for Group 1, with a CAT-Real value of 0.284 and a CAT-Sim value of 0.355 resulting in a difference of 0.071. The CAT-Real condition had an overall mean of 0.676 with a CAT-Sim value a bit larger at 0.691, for a total overall difference of 0.015.

Real-Simulated Information Summary. Comparing real and simulated data conditions, the CAT-Sim data resulted in the largest overall mean item information, with the 25-item CAT-Sim condition attaining the highest overall mean across all conditions at 0.776. The lowest mean item information was found for 25-item Linear-Real condition at 0.238. The highest group mean information was found for Group 4 across all test lengths and conditions. Across data conditions the highest information was found in the middle to lower proficiency groups, with neither simulated nor real data providing notably more mean information than the other. Test delivery format has a much larger impact on information values than does the use of real or simulated data, which were largely quite similar to each other for these information results.

4.2.7 Relative Efficiency

In comparing real to simulated data, it is often instructive to ascertain the level of relative efficiency, a statistic derived from item information as described in the previous chapter. By design, computer-adaptive tests are expected to be more efficient at measuring proficiency using fewer items than a linear test, while concurrently maintaining the same or lower standard error. Thus, comparisons between linear and CAT formats will not be presented in tabular format here. Summarizing the linear to CAT efficiency values, the smallest CAT value of 2.42 was found for the 25-item real data condition. The largest CAT-linear efficiency value of 3.13 was found for the 35-item test using simulated data. Table 4.19 summarizes the relative efficiency of real over simulated data (as per the formula from Chapter 3) within each test format and each test length condition within the two formats. Values greater than 1.0 in the matrix indicate that the real data is more efficient than the simulated data. Values less than 1.0 indicate that the real data created less efficient tests than did its simulated data counterpart. Generally, the overall efficiency measures are quite similar with values near 1.000 for most test lengths in both linear and CAT formats. The 35-item linear test overall efficiency is the smallest value and the only one notably below 1.000 at 0.985. The 50-item CAT format overall efficiency was the largest across all conditions at 1.018, indicating improved efficiency for real data over simulated data. For the linear test format, the group with the smallest efficiency is Group 1 in the 50-item test length at 1.114. The smallest linear format efficiency value was also found for Group 1 in the 50-item test length at 0.862. Within the CAT format, the largest efficiency group value was found for Group 6 within the 50-item test length at 1.092. The smallest CAT group value was found for Group 1 in the 35-item condition at 0.566, by far the lowest value in the table. The smallest overall efficiency difference between linear and CAT formats was found for the 25-item test length at 0.006. Similarly, Group 5 within the 25-item condition had the smallest linear-to-CAT group difference of -0.004. The largest overall efficiency difference was found for the 35-item test at -0.019. The largest group linear-to-CAT difference was found for Group 1 in the 35-item condition at 0.388, a notably large group difference that far exceeded the next largest group difference of 0.211 for Group 1 in the 25-item condition. In summary, the overall efficiency values indicate that the real data were slightly more efficient or as efficient as the simulated data except for the 35-item linear test condition where the simulated data were more efficient. The largest group efficiency values were found for the 50-item condition in both linear (Group 4) and CAT (Group 6) formats. The smallest within-format values were found in Group 1 in both test formats. The 35-item Group 1 showed the largest efficiency difference between linear and CAT formats. Overall, however, only the 50-item CAT showed a small notable efficiency improvement for real data. Generally, the real to simulated data efficiency values were small with only notably small values for CAT format in Group 1. The conclusion here is that, while CAT offers major efficiency increases over linear data, when real and simulated data were compared for efficiency, the differences were small but most noteworthy in the extreme proficiency groups.

Table 4.19

Relative Efficiency for Real by Simulated Data

	Linear			CAT		
Group	25-Item	35-Item	50-Item	25-Item	35-Item	50-Item
1	0.941	0.953	0.862	0.730	0.566	0.800
2	0.981	0.929	0.998	0.964	0.991	1.011
3	1.016	0.979	0.960	1.036	1.045	1.042
4	0.983	0.877	1.114	1.011	1.030	1.010
5	0.963	1.067	1.007	0.967	0.991	0.980
6	1.018	1.016	0.987	1.075	1.074	1.092
7	1.106	1.000	0.953	1.049	1.050	1.069
8	1.067	1.061	0.960	1.041	1.003	1.040
9	1.012	1.096	1.048	0.866	0.899	0.955
10	0.947	0.991	1.076	0.879	0.972	0.988
ALL	1.006	0.985	1.002	0.999	1.004	1.018

4.2.8 Correlation between True Proficiency and Estimated Proficiency

One effective method of summarizing the impact of research manipulations on recovery of true proficiency is the Pearson correlation. For this study, correlations for all conditions are summarized in Table 4.20. While prior tables have utilized stratified proficiency to summarize information, correlations are not necessarily amenable to this kind of analyses as a result of range restriction problems (as noted in the biserial correlation section). Therefore, only the overall summary correlations are reported in Table 4.20. The strongest correlation was exhibited by the CAT formats, with the highest correlations for simulated data, and lower correlations for the real data. Not surprisingly, the longest tests revealed the highest correlations, with the CAT-Sim data correlation value the highest at 0.983. The lowest true-to-estimated proficiency correlation of 0.923 was found for the 35-item Linear-Real condition.

Table 4.20

True Proficiency by Estimated Proficiency Correlations

	Linear Format		CAT Format	
Test Length	Real Data	Simulated Data	Real Data	Simulated Data
25	0.937	0.924	0.963	0.969
35	0.923	0.934	0.970	0.978
50	0.946	0.954	0.975	0.983

In comparing proficiency correlations across data types, simulation data were found to correlate more highly with true proficiency than real data. Within the linear test format, the difference between the real and simulated data was greater than within the CAT format conditions. The largest absolute difference of 0.013 was found for the 25-item linear test form. As linear test length increased, the differences between real and simulated data decreased. The differences within CAT forms were smaller than within linear forms, but the CAT differences increased slightly as test length increased. This outcome can be explained by noting that the correlations increased as test length increased, and the CAT format conditions showed larger correlation increases than the linear forms. As a result, the difference values increased because of the extra correlation gains made by the CAT format exams. Notably, the 25-item test linear format data indicated that the real data showed a higher correlation than the simulated data. Overall, the simulated data were found to have higher correlations with the true proficiency values regardless of test format, with the single exception of the 25-item linear test condition.

4.2.9 CAT Item Exposure

In summarizing any CAT procedure, it is often instructive to state the frequency with which items were selected by the CAT maximum information algorithm. Given the large number of item administrations from the item pool, a graphical illustration is the most parsimonious method for presenting these data. Figures 4.13, 4.14, and 4.15 show the ordered item exposures frequencies for the CAT tests. As illustrated by the steeply spiked graph, some items were chosen rarely while others were chosen quite frequently. Given that an unconstrained maximum information CAT typically uses its most informative items first, it is reasonable that some items would be selected frequently compared to less informative items which would be selected infrequently or never.

Figure 4.13

CAT Item Exposure for 25-Item Test

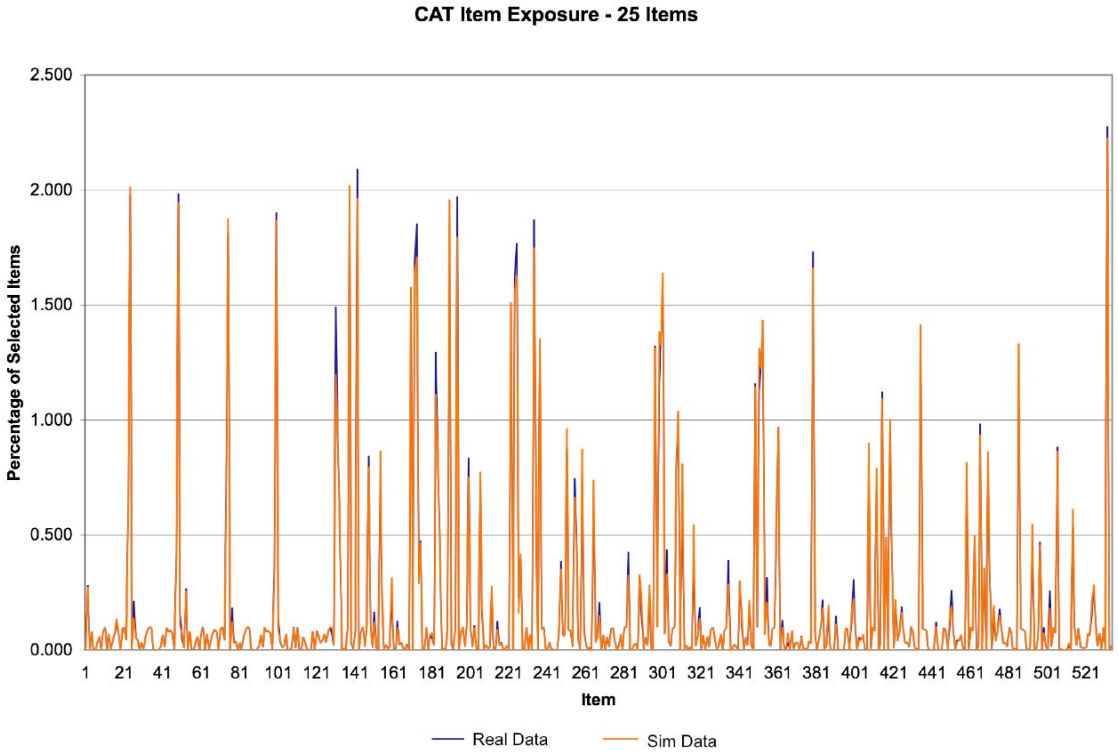


Figure 4.14

CAT Item Exposure for 35-Item Test

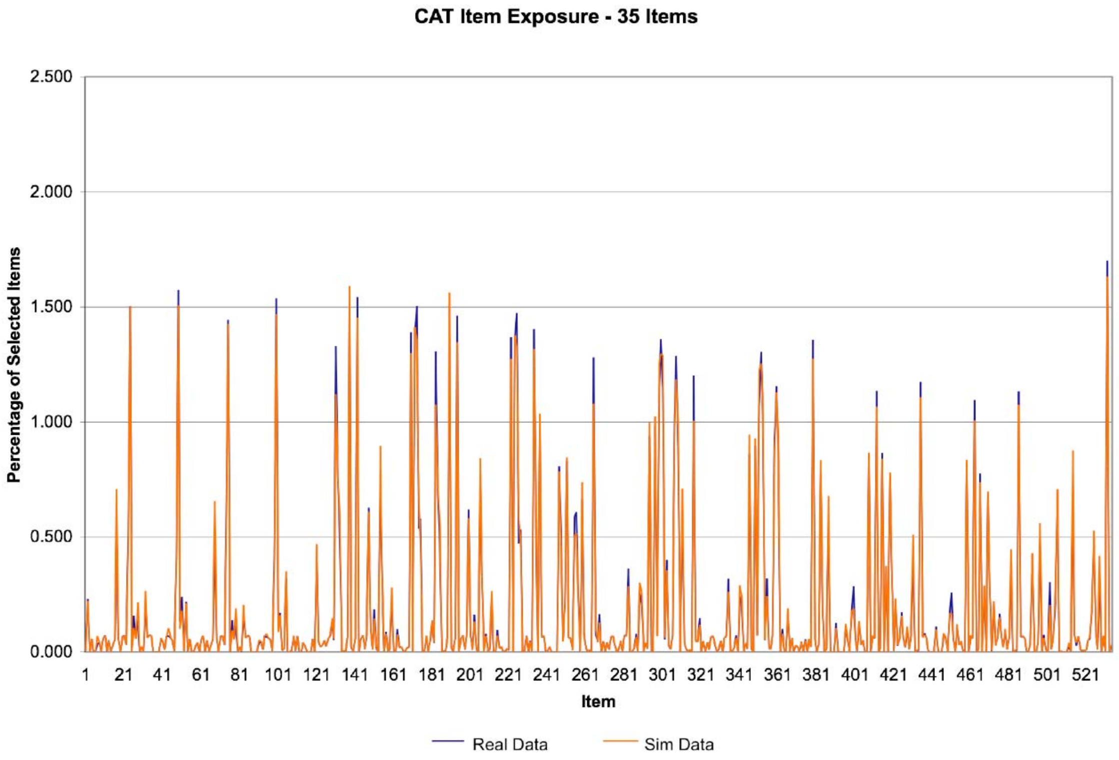
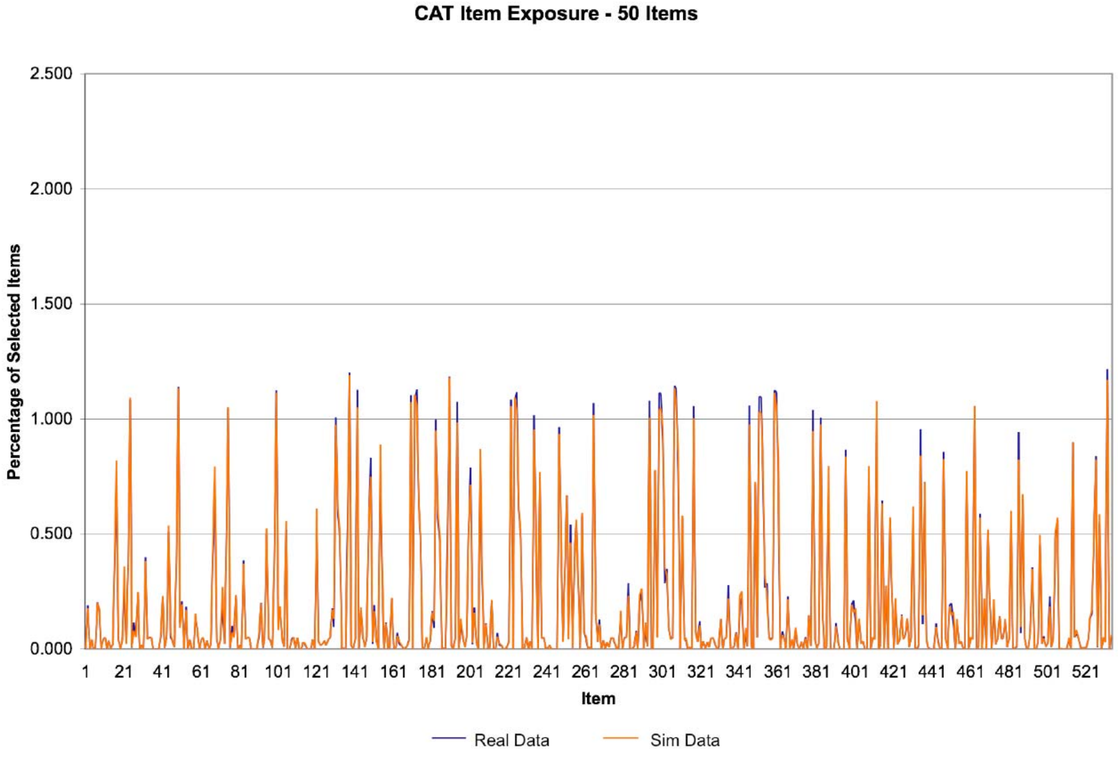


Figure 4.15

CAT Item Exposure for 50-Item Test



For all test lengths, the Figures show that generally the same items were chosen, regardless of whether or not the data were real or simulated. In some cases, certain items were chosen slightly more or less often depending on data type. For the 25-item test (Figure 4.13) and the 50-item test (Figure 4.15), the real data items were selected slightly more frequently except for a few points where they were chosen much more frequently. For the 35-item test, however, there were a larger number of items which were chosen much more frequently for the real data condition than for the simulated data condition. Again, this result is likely due to the reduced information in the 35-item condition, and the increased use of popular items by the CAT algorithm was an attempt to maximize information by using the most informative items. This tactic may have been less important for the simulated data which tended to have less variation than its real data counterpart.

Table 4.21

Item Exposure Summary for Real Data CAT

Test Length	Number of Items Used of 534	Number of Items Used ≥ 0.25%	Proportion of Linear Test Items Not Used by CAT
25	482	86	0.04
35	482	100	11.43
50	486	114	0.10

Table 4.22

Item Exposure Summary for Simulated Data CAT

Test Length	Number of Items Used of 534	Number of Items Used ≥ 0.25%	Proportion of Linear Test Items Not Used by CAT
25	484	85	0.04
35	483	102	11.43
50	484	114	0.10

Another point of interest is the number of items used from the total pool of 534 and a comparison of items selected. Tables 4.21 and 4.22 summarize the CAT item exposure. Items used for more than 0.25 percent of the available selections indicated that the items were used quite frequently. Of particular interest is the rightmost column which summarizes the number of items on the linear forms not chosen for any CAT administrations. For the 35-item test, four of the items used on the less stable linear test forms were never administered using the CAT algorithm. Therefore, more than 11% of the 35 linear test items were deemed unsuitable for administration by the CAT maximum information algorithm. This result confirms that the 35-item linear test was constructed with items that would not likely be included on a real-world test due to their poor performance, which is a drawback of the random selection procedure employed for the linear test assembly.

4.3 Calculated MSE Compared to Empirical SD for Simulated Data

As noted in the methodology chapter, the standard error of the proficiency estimate, $SE(\hat{\theta})$, is a function of the amount of information provided by a given test at $\hat{\theta}$. The $SE(\hat{\theta})$ statistic represents the amount of error in the proficiency estimate. Unlike the classical test theory standard error of measurement, which is the same value for all examinees on a given test, the $SE(\hat{\theta})$ is specific to the examinee's proficiency estimate. The *SE* is intended to represent the amount of potential within-examinee variation in the proficiency estimate that would occur over multiple administrations of the test, represented here as the "Mean $SE(\hat{\theta})$" or $\overline{SE}$. Given the 100 simulated replications for each test length condition in this study, it is possible to obtain the empirical standard deviation of the proficiency estimate, $\mathrm{SD}\left(\hat{\theta}\right)$, at the examinee level. This $\mathrm{SD}\left(\hat{\theta}\right)$ value, or SD, can be compared to the $\overline{SE}$ to show differences in proficiency estimation by method of test administration (i.e., CAT versus linear conditions). Comparisons of $\overline{SE}$ and $\mathrm{SD}\left(\hat{\theta}\right)$ values for simulated data on the linear test are summarized in Table 4.23. Table 4.24 summarizes the same comparisons for the simulated data CAT tests. On the linear tests (Table 4.23), the overall $\overline{SE}$ values were larger than the overall SD values with the 25-item test having the largest overall $\overline{SE}$-SD difference of 0.039. The smallest overall difference was 0.019 for the 50-item test. Among the groups, the same pattern holds true, except for the middle of the proficiency distribution, Group 4 and Group 5.

Table 4.23

Linear Test Simulated Data $\overline{SE}$ *Compared to* $SD\left(\hat{\theta}\right)$

	$\overline{SE}$			$SD\left(\hat{\theta}\right)$		
Group	25 Item	35 Item	50 Item	25 Item	35 Item	50 Item
1	0.430	0.373	0.321	0.335	0.309	0.274
2	0.338	0.290	0.233	0.320	0.281	0.231
3	0.323	0.280	0.223	0.316	0.274	0.223
4	0.307	0.279	0.223	0.312	0.293	0.232
5	0.310	0.286	0.231	0.311	0.297	0.238
6	0.336	0.329	0.271	0.325	0.315	0.266
7	0.341	0.335	0.277	0.328	0.316	0.270
8	0.391	0.389	0.329	0.356	0.340	0.303
9	0.419	0.415	0.355	0.362	0.349	0.318
10	0.492	0.482	0.423	0.340	0.356	0.341
ALL	0.369	0.346	0.289	0.330	0.313	0.270

Across all test lengths, these two groups only had larger $\overline{SE}$ values than SD values. The group with the largest $\overline{SE}$ -SD difference was Group 10, across all three test lengths. Among the three test lengths, the largest Group 10 difference was found for the 25-item test at 0.152 and the smallest Group 10 difference was found for the 50-item test at 0.082. Within the groups, the smallest $\overline{SE}$ -SD difference was found for Group 3 in both the 35-item test (0.006) and 50-item test (0.000). The second largest difference for the 25- and 50-item tests were in Group 1 for both, but the 35-item test had its second largest difference in Group 9. For the 25-item test, the smallest difference was found for Group

5 at -0.001. Overall, the linear simulated data showed that $\overline{SE}$ values were larger than SD values overall and for all groups except Groups 4 and 5. Linear-Sim $\overline{SE}$ -SD group differences were largest for Group 10 and smallest for Group 3 (35- and 50-item tests) and Group 5 (25-item tests). The largest overall differences were found for the 25-item test and the smallest were found for the 50-item test.

Table 4.24

CAT Test Simulated Data $\overline{SE}$ *Compared to* $SD\left(\hat{\theta}\right)$

	$\overline{SE}$			$SD\left(\hat{\theta}\right)$		
Group	25 Item	35 Item	50 Item	25 Item	35 Item	50 Item
1	0.345	0.281	0.235	0.371	0.295	0.245
2	0.165	0.144	0.126	0.177	0.149	0.128
3	0.154	0.135	0.119	0.164	0.140	0.121
4	0.147	0.130	0.116	0.153	0.133	0.118
5	0.150	0.133	0.119	0.156	0.136	0.120
6	0.175	0.156	0.139	0.180	0.158	0.140
7	0.181	0.161	0.143	0.187	0.164	0.146
8	0.243	0.212	0.185	0.259	0.221	0.190
9	0.287	0.248	0.214	0.303	0.258	0.219
10	0.423	0.358	0.306	0.444	0.372	0.314
ALL	0.227	0.196	0.170	0.239	0.203	0.174

For the CAT format simulated data (Table 4.24), the SD values were larger than the $\overline{SE}$ values, in contrast to the results from the linear format data. However, the differences between the two statistics in the CAT data were smaller than they were for the linear data. Among the three test lengths, the largest difference value was found for the 25-item test at -0.012, and the smallest difference value was found for the 50-item test at -0.004. The largest group differences were found in Group 1 for two of the test lengths, the 25-item (-0.026) and the 50-item (-0.010) tests. The largest 35-item test group difference was tied for both Group 1 and Group 10 (-0.014). The group with the smallest difference was Group 6 for two test lengths, the 25-item (-0.005) and the 35-item (-0.002) tests. The 50-item (-0.001) test had its smallest group difference values for Group 5 and Group 6, tied at -0.001.

The larger $\overline{SE}$ values for the linear format tests indicate that the calculated errors are larger than they would be, given the smaller empirical SD. Thus, for most of the groups and for the overall linear test means, the $\overline{SE}$ overestimated the error compared to results found empirically. For the two middle linear groups that had negative difference scores, the SD was larger than the $\overline{SE}$ which signifies that the actual empirical errors were larger than their calculated errors. For these two groups and for all of the CAT format results, which also had larger SD than $\overline{SE}$ values, the $\overline{SE}$ underestimated the error as compared to the empirically derived SD. For both linear and CAT format tests, either the overestimation or underestimation results were greater for the shorter length tests and less for the longer tests. Linear tests also had notably larger $\overline{SE}$ -SD differences than did CAT tests. The largest group differences across all formats and test lengths were found in the extreme high and low proficiency groups, Group 1 and Group 10. The smallest difference values occurred in the middle proficiency groups, but no one group across all conditions.

4.4 Summary Across Analyses

Generally, what was found in this study varied across analytical methods and by condition, as would be expected in any study. There are some common findings, however, that may prove useful to reiterate. The goal of this study was to determine if real and simulated data were functionally equivalent, using a new method of joining real test data from one section of the LSAT. An additional focus was to determine if differences could be introduced by varying linear versus CAT examination methods as well as by varying three test lengths. Overall, this study found that real and simulated data were largely the same in outcomes from the various conditions, when compared within condition. Simulated data generally displayed less variation than real data, particularly for the middle proficiency examinees and longer test lengths. Broadly speaking, the 50-item test length minimized any small differences between real and simulated data. The poorly constructed 35-item Linear-Real test, however, showed the greatest amount of unpredictability, which serves as a reminder of the importance of careful test assembly.

As would be expected, the greatest impact on bias, RMSE, information, and MSE was found for the CAT testing format. All of the positive reasons for using a CAT were reinforced by the results from this study. Compared to its linear counterparts, the computer-adaptive testing method used here created small values for bias, RMSE, MSE, and empirical standard deviation as well as more information. These results would be expected given the information maximization algorithm of the CAT and information's direct relationship to estimation accuracy, error and proficiency estimation variation. The CAT format examinations in this study, however, heavily utilized some of the most informative items from the item pool, a potential liability in real-world testing environments where quality item writing can be an expensive endeavor. Heavily exposed items may not be usable over time, as they will become too memorable, possibly introducing construct irrelevant changes in item performance.

Overall, however, this study failed to find notable differences between real and simulated data when compared within matched conditions (e.g., within CAT or within the same test length only). A few results do suggest caution when creating quality linear tests and shorter-length tests which may impact simulated data results. Finally, these findings suggest that simulated data may be somewhat less accurate at reproducing real data characteristics for extremely low or extremely high proficiency examinees.

CHAPTER V. DISCUSSION

In measurement research, the ubiquity of data simulation studies could cause some discomfort from research consumers who may be making high stakes testing decisions based on the results of these artificially generated data. Empirically demonstrating the validity of using simulated data to inform measurement research choices is an important undertaking to ensure continued usefulness and acceptance of these methods. Few studies have examined the empirical basis of the validity of simulated data within the modern CAT format. Even fewer studies have compared these simulated data CAT results to more traditional linear test results using both real and simulated data. The purpose of this dissertation was to fill this void by examining one specific section of LSAT data to determine the comparability of simulated data to real-world examinee data.

Simulating data serves a valuable function by enabling educational measurement researchers to obtain timely answers to difficult research questions. In this dissertation, the design focused on methods establishing if simulations are sufficiently representative of real test data. Harwell et al. (1996) noted that well-designed simulation studies can be very effective and accurate in creating realistic data. Using Harwell et al.’s recommendations regarding careful construction of the simulated data, support for the accuracy of simulated data in this study was high. Unfortunately, this study’s results did not indicate a perfect relationship between simulated and real examination data. Reduced simulation accuracy was found on the linear test formats by way of increased real-to-simulated data differences in proficiency estimates, true-by-estimated correlations, bias, and RMSE. In particular, the problematic real data linear test with 35 items was shown to have nontrivial differences as compared to the simulated data. The item selection method for the linear tests was simple stratified random sampling, which had no constraints or item rejection criteria of any kind. Unfortunately, this random selection process resulted in the inclusion of poorly performing items which may have exacerbated differences between the real and simulated data. Since the CAT method selects items that are appropriate for the proficiency-derived response pattern, items chosen were properly targeted to each examinee. Therefore, CAT format data showed increased accuracy between real and simulated data.

On the linear tests, carefully selected item sets may show better recovery of proficiency estimates, as shown with the 25-item real data condition. Within that condition, bias, RMSE, and correlations with true proficiency indicate that the real data captured proficiency more effectively than the simulated data. On the 35-item linear test which was composed of items having less ideal item parameters, the simulated data were somewhat less extreme than the more atypical results found in the real data. The simulated data behaved more as it would be expected to behave as dictated by theory, with smaller values for bias and RMSE, and an increased true proficiency correlation over the shorter 25-item simulated data test. Since CATs select targeted, highly informative items, inefficient item selection did not negatively impact results. Thus, simulated data may recover proficiency estimates more accurately in a CAT format because all data types recover proficiency estimates more accurately when optimally informative items are selected. Moreover, the alignment between CAT and simulated data is high, as both were expressly created to capitalize on features of the IRT theoretical model.

Limitations and Future Research

Every study has its limitations and this dissertation is no exception. The creation of the synthetic examinee, post-calibration item fit assessment, and linear item selection constraints are some of the limitations to this research project.

Combining and matching real item response vectors using the complex method presented in this paper is untested within the measurement research field. The potential benefits of having a large item pool of real examinee responses, however, are of sufficient value to outweigh potential unease with using this new method. Many advances have been made in the area of string comparisons of many types and some advanced applications in other fields may be of use in refining this process (e.g., Sankoff and Kruskall, 1999). The notable depression in the c-parameter estimates indicates that the synthesizing process might have actually altered the original data, perhaps by replicating and thereby exacerbating the existing error variance from the original data. Thus, the synthesized data would have benefited from some method to reduce the compounded error. Regardless of the method used to reduce error, documenting any differences between the original data and the synthesized data would be instructive. Additional validation research on this particular aspect would enable future researchers to refine the practice so that a reliable standard can be achieved.

Items chosen for use in the study were also a limiting factor, likely compounded by the aforementioned non-normal distribution of the final matrix. In this study, the measure of an item’s suitability for inclusion on a linear test was simple: If it converged in the calibration and did not display any extremely unusual characteristics, it was deemed a suitable item. Unfortunately, this measure of an item’s fitness is insufficient for the purpose of including it on a real test. Items with extreme item parameters are typically evaluated in multiple ways (e.g., stats, content, administration errors) before inclusion in a real-world item pool. Most such items are pre-tested as unscored items and then retested if there is a question as to the item's suitability. In this project, the focus was on creating randomly selected item groups to avoid any assembly biases and to maintain reasonable parsimony. Unfortunately, the random item selection process chose several items with more extreme values for the linear 35-item test. A more tightly controlled item selection process may be more effective for evaluating future simulation data research questions. Moreover, the lack of comparability between the linear tests in this study would preclude direct comparison, a limitation that could have been avoided by constructing appropriately parallel tests (as in Sanders & Verschoor, 1998). In this study, a potentially improved method of constructing tests would be to restrict the item pool to only items which have parameters within a particular ideal range. Stratifying and selecting items only after such item filtering would reduce the possibility of creating a test with outlying item parameters.

In addition to the issue of item selection based on statistical properties, the lack of real-world item selection constraints is a potential limitation in this study. Without selection constraints, such as content constraints, one could argue that the study fails to capture truly realistic testing conditions. While this point is a valid criticism, the main point of this research study was not to perfectly mimic a realistic testing scenario with the full complement of testing conditions. In fact, some real-world item constraints can be exceedingly complex, requiring professional optimization software to resolve all the selection restrictions (van der Linden, 1998). The focus of this research paper was to compare results of using real-world responses to those of simulated responses from a measurement perspective. Additional foci were test format and number of items administered. A key point is that this is a baseline study and the intent was to limit the scope to a reasonable level. To address more complex questions, future research can incorporate some of the complex optimization issues arising from multiple constraints.

For the current study, some specific improvements could help to refine the results. One improvement would be to set simple conditions for the linear test item selection process. The current methodology allowed items to be chosen at random from within strata, which is a method that, while parsimonious, is not particularly realistic. It is highly unlikely that any modern test developer would allow a computer to randomly select items from an unfiltered item pool, even if the items are stratified to sample the most informative items for all parts of the ability continuum. Items with more extreme parameter estimates would not likely be chosen in a real-world situation. By removing items with more extreme parameters, a selection algorithm could choose from among a more informative item set. In turn, the real and simulated data should be more closely aligned. Alternatively, setting up various linear test item combinations may prove informative, such that the more poorly assembled tests could be treated as an independent variable to be manipulated. That sort of study would elucidate at what point simulated and real data diverge and to what extent that divergence impacts important dependent variables such as proficiency estimates.

Besides item selection changes, this study could have benefited from additional improvements in response modeling manipulations. The 3PL IRT model was the only model tested in this study. It may have been informative to examine additional models which may be more capable of illustrating differences between real and simulated data. For example, perhaps the use of a more multivariate-type model, or one including other cognitive factors would have been beneficial. Another improvement would have been to develop a better way to create the synthetic examinee. As noted previously, the method used here is untested and therefore may not be the ideal method to use to create a synthetic examinee. Some highly advanced methods of matching data strings exist in the realm of physics and biology, from bird songs to DNA matching. These methods, while complex, are sophisticated and may provide a better match to create the synthetic examinee.

Some methods that can be used to compare simulated data to the real counterpart were explored here, and may be useful to researchers attempting to ascertain the comparability of their simulated data to a real data counterpart. One example is the comparison of proficiency estimates and standard errors within stratified groups between the two data types. Another method that may prove useful is comparing bias and RMSE statistics as well as the item information differences between the two data types. Correlations between the real data proficiency estimates and simulated data proficiency estimates may also be a useful method of demonstrating comparability. Of particular interest, researchers may want to focus particularly on the lowest and highest proficiency values of their simulated data, as the largest real-to-simulated data differences in this study were found for those parts of the distributions.

The implications of this study on in the field of educational measurement are preliminary, but promising. As noted in the literature review, some published research has sounded the alarm regarding the use of poorly constructed simulated data. Many of these publications note of the options for improving simulated data so that they better reflect real world data. These publications, however, have been unsatisfactory for the various reasons outlined previously. The implications for this study are the plausibility of using simulated data, with caveats, and the usefulness of creating the synthetic examinee. Simulated data in this study were generally quite similar to their real data counterparts, at least when the tests constructed were parallel and of sufficient length. As noted, the simulation of highest and lowest proficiency groups was the most problematic. The implication of these findings is that well constructed simulations are a viable method of creating data for use in psychometric analyses. The inherent difficulty in constructing a valid synthetic examinee may be one hindrance to a viable solution to this issue. This study outlines a preliminary method that may prove useful as a new method of comparing simulated data to real data by way of creating the synthetic examinee dataset. By refining this synthetic examinee process, it is hoped that this method will prove useful to researchers seeking to compare their simulated datasets to real test data. With additional research on the validity of the synthetic examinee output, it may prove to be a very useful method indeed.

References

Ansley, T. N. & Forsyth, R. A. (1985). An examination of the characteristics of the unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement, 9* (1), 37-48.

Ban, J.-C., Hanson, B. A., Yi, Q., Harris, D. J. (2002). *Data sparseness and online pretest item calibration/scaling methods in CAT.* ACT Research Report Series. Iowa City, IA: American College Testing Program.

Bobko, P. (1983). An analysis of correlations corrected for attenuation and range restriction. *Journal of Applied Psychology,* 68 (4), 584-589.

Bowles, R. & Pommerich, M. (2001, April). *An examination of item review on a CAT using the Specific Information Item Selection algorithm.* Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Chang, S.-W. & Twu, B.-Y. (2001, April). *Effects of changes in the examinees' ability distribution on the exposure control methods in CAT.* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Harcourt, Brace, & Jovanovich.

Davey, T., Nering, M. L. & Thompson, T. (1997). *Realistic simulation of item response data*. ACT Research Report Series, 97-4, American College Testing, Iowa City, IA.

De Ayala, R. J. (1994). The influence of multidimensionality on the graded response model. *Applied Psychological Measurement, 18* (2), 155-170.

De Boeck, P. & Wilson, W. (Eds.). (2004). Explanatory item response models: A generalized linear and nonlinear approach. New York: Springer-Verlag.

de la Torre, J. (2009). A Cognitive Diagnosis Model for Cognitively Based Multiple-Choice Options. *Applied Psychological Measurement, 33* (3), 163-189.

de la Torre, J. & Douglas, J. A. (2008) Model Evaluation and Multiple Strategies in Cognitive Diagnosis: An Analysis of Fraction Subtraction Data. *Psychometrika, 73* (4), 595-624.

Evans, J. and Weissman, A. (2005, October). *IRT 3PL Parameter Recovery under Sparse Data Conditions.* Paper presented at the annual meeting of the Northeast Educational Research Association, Kerhonkson, NY.

Fairbank, B. A., Jr. (1985, April). *Equipercentile test equating: The effects of presmoothing and postsmoothing on the magnitude of sample dependent-errors.* (Research report AFHRL-TR-84-64). San Antonio, TX: Performance Metrics, Inc.

Fan, X., Felsovalyi, A., Sivo, S.A., and Keenan, S.C. (2001). *SAS® for Monte Carlo studies.* Cary, NC: SAS Institute.

French, J. W. (1965). The relationship of problem-solving styles to the factor composition of tests. *Educational and Psychological Measurement, 25*, 9-28.

Hambleton, R. K. (1993). Principles and selected applications of item response theory. In R. L. Linn (Ed.) *Educational measurement* (3rd ed.). Phoenix, AZ: American Council on Education / Oryx Press.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Harwell, M. R., & Janosky, J. E. (1991). An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in BILOG. *Applied Psychological Measurement, 15*, 279-291.

Harwell, M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*, 101-125.

Hoaglin, D. C., & Andrews, D. F. (1975). The reporting of computation-based results in statistics. *American Statistician, 29*, 122-126.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement, 6*, 249-260.

Kim, H., & Plake, B. S. (1993, April). *Monte Carlo simulation comparison of two-stage testing and computerized adaptive testing*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Atlanta, GA.

Kubinger, K. D. (2009). Applications of the linear logistic test model in psychometric research. *Educational and Psychological Measurement, 69* (2), 232-234.

Levine, M. V. (1984). *An introduction to multilinear formula score theory*. Model-Based Measurement Laboratory Report 84-4. Urbana: University of Illinois.

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

Lord, F. M. & Novick, M. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

McCauley, C. D., & Mendoza, J. (1985). A simulation study of item bias using a two-parameter item response model. *Applied Psychological Measurement, 9*, 389-400.

McLeod, L. D., Lewis, C., & Thissen, D. (1999). *A Bayesian method for the detection of item preknowledge in CAT.* LSAC Computerized Testing Report. Newtown, PA: Law School Admission Council.

Metropolis, N. & Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association, 44*, 335-341.

Mills, C. N. & Stocking, M. L. (1996). Practical issues in large-scale computer-adaptive testing. *Applied Measurement in Education, 9* (4), 287-304.

Mooney, C. Z. (1997). *Monte Carlo simulation.* Thousand Oaks, CA: Sage.

Parshall, C.G., Spray, J.A., Kalohn, J., Davey, T. (2002). *Practical considerations in computer-based testing.* New York: Springer.

Prowker, A. N. (2005). Long-term stability of fixed common item parameter equating: What No Child Left Behind could mean for equating practices. (Doctoral dissertation, Rutgers, The State University of New Jersey - New Brunswick, 2005). *Dissertation Abstracts International, 66* (11). (Proquest Publication No. AAT 3195741)

Psychometric Society (1979). Publication policy regarding Monte Carlo studies. *Psychometrika*, *44*, 133-134.

Sanders, P. F. & Verschoor, A. J. (1998). Parallel test construction using classical item parameters. *Applied Psychological Measurement*, *22*, 212-223.

Sankoff, D. & Kruskall, J. (1999). *Time warps, string edits, and macromolecules: The theory and practice of sequence comparison.* Stanford, CA: Center for the Study of Language and Information (CLSI) Publications.

Schnipke, D., Roussos, L., & Pashley, P. (2000). *A comparison of Mantel-Haenzel differential item functioning parameters* (Law School Admission Council Research Report No. RR-98-03). Newtown, PA: Law School Admission Council.

Snow, R. E. & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In Linn, Robert L. (Ed). *Educational measurement* (3rd ed.). New York: The American Council on Education and Macmillan Publishing.

Sternberg, R. J. & Weil, E. M. (1980). An aptitude-strategy interaction in linear syllogistic reasoning. *Journal of Educational Psychology, 72* (2), 226-234.

Stocking, M. L., Steffen, M., & Eignor, D. R. (2001). *A method for building a realistic model of test taker behavior for computerized adaptive testing* (Educational Testing Service Research Report, RR-01-22). Princeton, NJ: Educational Testing Service.

Thissen, D. & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 101-134). Mahwah, NJ: Lawrence Erlbaum Associates.

van der Linden, W. (1998). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement*, *22*, 195-211.

Wainer, H. (2000). *Computer adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Wainer, H. & Mislevy, R. J. (2000). Item response theory, calibration, and estimation. . In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 61-100). Mahwah, NJ: Lawrence Erlbaum Associates.

Wang, X. B., Pan, W., & Harris, V. (1999). *Computerized adaptive testing simulations using real test taker responses* (Law School Admission Council Computerized Testing Report, 96-06). Newtown, PA: Law School Admission Council.

Weiss, D. J. (2005a). *Manual for POSTSIM: Post-hoc simulation of computerized adaptive testing. Version 2.0.* St. Paul, MN: Assessment Systems Corporation.

Weiss, D. J. (2005b). POSTSIM 2.0 [computer software]. St. Paul, MN: Assessment Systems Corporation.

Wen, J.-B., Chang, H.-H., & Hau, K.-T. (2000, April). *Adaptation of a stratified method in variable length computerized adaptive testing.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA, 2002.

Yen, W. M. (1987). A comparison of the efficacy and precision of BILOG and LOGIST. *Psychometrika, 52*, 275-291.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R.D. (2003). BILOG-MG 3 for Windows: Multiple-group IRT analysis and test maintenance for binary items [Computer software]. Lincolnwood, IL: Scientific Software International, Inc

Zwick, R. & Thayer, D. T. (2003). *An empirical Bayes enhancement of Mantel-Haenzel DIF analysis for computer adaptive tests* (LSAC Research Report). Newtown, PA: Law School Admission Council.

Curriculum Vitae

Josiah Evans

Formal Education

Master of Arts, Psychology, May 2000
Hunter College/The Graduate Center – City University of New York, New York, NY

Bachelor of Science, Psychology, May 1994
Presbyterian College, Clinton, SC

Relevant Employment Experience

Research Associate 2003 – present
Law School Admission Council, Newtown, PA

Assistant Psychometrician 2000 – 2003
Research Assistant 1999 – 2000
The American Institute of Certified Public Accountants, Jersey City, NJ

Research Assistant 1997-1999
Mount Sinai School of Medicine, New York, NY

Research Assistant 1996-1997
GMHC, New York, NY

Publications

Evans, J., Thornton, A. E., & Reese, L. M. (2008) *Summary of self-reported methods of test preparation by LSAT takers for testing years 2005–2006 through 2007–2008* (Law School Admission Council Technical Report 08-04). Newtown, PA: Law School Admission Council.

Mills, C., Hambleton, R., Biskin, B., Kobrin, J., Evans, J., and Pfeffer, M. (2000). *Setting passing standards using two methods: Cluster and Angoff.* Technical report prepared for the National Association of the State Boards of Accountancy, Board of Examiners, Nashville, TN.