University of Massachusetts Amherst
ScholarWorks@UMass Amherst

Doctoral Dissertations | Dissertations and Theses

April 2014

Effect of Automatic Item Generation on Ability Estimates in a Multistage Test

Kimberly F. Colvin
University of Massachusetts Amherst

Follow this and additional works at: https://scholarworks.umass.edu/dissertations_2

 Part of the Educational Methods Commons, and the Other Education Commons

Recommended Citation

Colvin, Kimberly F., "Effect of Automatic Item Generation on Ability Estimates in a Multistage Test" (2014). *Doctoral Dissertations*. 4.
https://doi.org/10.7275/5428533 https://scholarworks.umass.edu/dissertations_2/4

This Open Access Dissertation is brought to you for free and open access by the Dissertations and Theses at ScholarWorks@UMass Amherst. It has been accepted for inclusion in Doctoral Dissertations by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.

EFFECT OF AUTOMATIC ITEM GENERATION ON ABILITY ESTIMATES IN A MULTISTAGE TEST

A Dissertation Presented

by

KIMBERLY F. COLVIN

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF EDUCATION

February 2014

Education
Psychometric Methods, Educational Statistics and Research Methods

© Copyright by Kimberly F. Colvin 2014

All Rights Reserved

EFFECT OF AUTOMATIC ITEM GENERATION ON ABILITY ESTIMATES IN A MULTISTAGE TEST

A Dissertation Presented

by

KIMBERLY F. COLVIN

Approved as to style and content by:

_____
Lisa A. Keller, Chair

_____
Erin M. Conlon, Member

_____
Craig S. Wells, Member

_____
Christine B. McCormick
Dean of the College of Education

DEDICATION

To my husband, Ron, without whom, none of this would have been possible.

ACKNOWLEDGMENTS

I was introduced to automatic generation by my mentor Isaac Bejar during a summer internship at Educational Testing Service (ETS). While there, my discussions with both Isaac and Frederic Robin led me to this dissertation topic. I would like to thank ETS for its financial support and Fred for his direction during my dissertation work.

The former and current students of the University of Massachusetts' Psychometrics Program, formally known as REMP, have been supportive, strong colleagues and wonderful references during my doctoral studies. It is obvious from my interactions with graduates of this program that the strong connections among "REMPers" will continue throughout our careers. These strong relationships are fostered by Professors Ronald Hambleton and Stephen Sireci, who are wonderful at connecting current and former students and other members of the psychometric community.

I am grateful to Professor Jennifer Randall, a fellow, former high school teacher, for allowing me to observe her course on classroom assessment. Apart from my advisor, I spent the most time working with Jennifer on various projects and I am appreciative of her critiques of my writing on the several literature reviews we completed together. Even though I learned a great deal from Professor Craig Wells' other courses, the R programming course was invaluable to my dissertation work and every other project I have worked on since then. His guidance, along with that of Professor Erin Conlon, was invaluable during this process.

Finally, in my advisor, Professor Lisa Keller, I found someone with a similar background who was able to offer such wonderful advice and guidance. She understands my strengths and weaknesses and has helped me to work on the latter and take advantage of the former. I know I will be relying on Lisa for both her professional advice and her friendship for years to come.

ABSTRACT

EFFECT OF AUTOMATIC ITEM GENERATION ON ABILITY ESTIMATES IN A MULTISTAGE TEST

FEBRUARY 2014

KIMBERLY F. COLVIN, B.S., CORNELL UNIVERSITY

M.A.T., CORNELL UNIVERSITY

M.A., UNIVERSITY OF ROCHESTER

Ed.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Lisa A. Keller

In adaptive testing, including multistage adaptive testing (MST), the psychometric properties of the test items are needed to route the examinees through the test. However, if testing programs use items which are automatically generated at the time of administration there is no opportunity to calibrate the items therefore the items' psychometric properties need to be predicted. This simulation study evaluates the accuracy with which examinees' abilities can be estimated when automatically generated items, specifically, item clones, are used in MSTs. The behavior of the clones in this study was modeled according to the results of Sinharay and Johnson's (2008) investigation into item clones that were administered in an experimental section of the Graduate Record Examination (GRE). In the current study, as more clones were incorporated or when the clones varied greatly from the parent items, the examinees' abilities were not as accurately estimated. However, there were a number of promising conditions; for example, on a 600-point scale, the absolute bias was less than 10 points for most examinees when all items were simulated to be clones with small variation from their parent items or when all first stage items were simulated to have moderate variation from their parents and no items in the second stage were cloned items.

TABLE OF CONTENTS

Page

APPENDICES

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background

The 20th century was witness to incredible developments in technology that have permeated almost every aspect of our lives, including transportation, manufacturing, food production, entertainment, and how we communicate with each other, to name just a few. Specifically, the improvements in and prevalence of computing capabilities can be seen in both education and in assessment. According to the National Center for Education Statistics, in 2009 97% of teachers had at least one computer in their classroom and 93% had access to the Internet (U. S. Department of Education, 2010). Many teachers use technology as a matter of course in their instruction. Schools are only now beginning to capitalize on the power of technology in the classroom. Some districts use individual tutorials that can be used to identify weaknesses, reinforce skills, or simply allow students to work at their own pace through material tailored to their own abilities.

The advances in technology have also spread to assessment, notably in the realm of adaptive testing. Adaptive testing refers to a test in which each item or set of items, referred to as item-level or multistage testing, respectively, is chosen to match the ability level of the examinee. While an adaptive test can be administered in the context of a paper-based test where the test administrator follows a prescribed set of rules for selecting the next item, adaptive testing is most easily administered on a computer. Adaptive testing is used in achievement, licensure, and credentialing examinations. The Graduate Management Admission Test (GMAT) uses an item-level computer adaptive test (Graduate Management Admission Council, n. d.) and both the Graduate Record Examination (GRE) and Uniform

CPA Exam employ multistage testing (Educational Testing Service, n. d., American Institute of Certified Public Accountants, n. d.). In K-12 assessment the two consortia, The Smarter Balanced Assessment Consortium (SBAC) and the Partnership for Assessment of Readiness for College and Careers (PARCC) intend to deliver computer-based assessments; SBAC plans to use an item-level computer adaptive test for both interim diagnostic testing and summative assessments (SBAC, n. d., PARCC, n. d.).

For a variety of reasons, testing programs often have a need for many test items. For one, test security is an issue for almost all testing programs. Programs like the GRE, GMAT, and Uniform CPA examinations offer their tests several times during the year therefore they need many items so that examinees will see different items during different administrations. Regardless of the need for more and more items, these testing programs which also employ computer-adaptive testing could benefit from supplementing their item pool with automatically generated items: New items created by altering some key features of a parent item. Before considering how they will interact, let us briefly consider the main ideas just mentioned: item-level computer adaptive testing, multistage adaptive testing, and automatic item generation.

1.2 Item-Level Computer Adaptive Testing

In item-level adaptive testing each item selected is intended to match the proficiency level of the examinee as estimated by the examinee's responses up to that point. This is an efficient testing procedure that can arrive at an examinee's estimated proficiency level within a specified degree of accuracy faster than a fixed-length linear test (Wainer, 1990). Some additional benefits of CAT are that the examinees who may normally feel frustration during traditional testing, will be less frustrated because the items will be at their ability level rather

than being too difficult or too easy (Lord, 1977); innovative item types that are difficult or even impossible to administer on a paper-based exam can be administered on a computer (Bennett, 2002). Examinees are unlikely to see the same items, so there is less of a security threat when examinees take the exams a few weeks apart, for example (Wainer, 1990).

However, there are some drawbacks to item-level adaptive testing. While many constraints, related to item format or content, can be placed on item selection algorithms, it is not always possible to meet all statistical requirements and other constraints. This could result in two examinees seeing different tests in terms of content specifications (Luecht & Nungester, 1998). In CATs, many students have expressed frustration that they cannot skip over or return to an item (Mead, 2006; Patsula, 1999; Vispoel, Rocklin, & Wang, 1994). Additionally, there are issues related to how long certain items take to complete. Items similar in quantity and difficulty can take different lengths of time to complete, even for examinees of the same ability level. This time difference can result in two equally able examinees receiving two exams equivalent in difficulty but quite different in length of time to complete (Bridgeman & Cline, 2004). This inequity can lead to a great disadvantage for some examinees. The benefits of CAT notwithstanding, the disadvantages pose a threat to the validity of the exams.

1.3 Multistage Testing

Multistage testing (MST) is a popular form of adaptive testing that retains most of the positive features of item-level adaptive testing and improves upon some of its less desirable characteristics. Multistage testing adapts the test to the examinee by sets of items, known as modules (Luecht & Nungester, 1998). An examinee responds to all items in a module and is then routed to the next module, rather than the routing occurring after each

item as in item-level CATs. Multistage testing can alleviate some of the issues resulting from item-level adaptive testing. Unlike item-level adaptive testing where the final test form is not even constructed until test administration, in MST the test developers can inspect all possible test forms that examinees could face. This allows content experts to examine the entire test form before administration to inspect for potential issues related to item content, item format, or issues that were impossible to write constraints for (Hendrickson, 2007). Even though the test developer may include many constraints in the item-level adaptive procedure, there is always the possibility that a potential issue or conflict will not be recognized until someone actually looks at the complete test form. From the examinee's point of view, MST can offer a reduction in test anxiety in that the examinee can move around within a module, can skip a question and return to it later (Hendrickson, 2007). Because the modules are assembled ahead of time, consideration can be given to the expected time to complete for each module. It is more likely that the test developers can design modules of 10 items, say, that are approximately equal in time to complete than multiple samples of any 10 items from an item pool are equal in time to complete (Bridgeman & Cline, 2004).

Multistage testing is less efficient in arriving at an examinee's proficiency estimate than an item-level CAT, but still more efficient than a fixed-length linear test (Hendrickson, 2007). The features of a multistage test such as the number of stages, the number of items per module, the number of modules per stage all affect the efficiency and accuracy of the resulting estimates (Zenisky, Hambleton, & Luecht, 2010). Another feature to consider is the possible paths that the examinees are allowed to take: Examinees could be allowed to jump from a hard to a medium module, but not to an easy module, for example (Luecht & Nungester, 1998). The flexibility in MST features coupled with the advantages of computer-

based testing has led to MST's increasing popularity. It can be used for both large-scale assessments and for use in formative assessments in the classroom.

1.4 Automatic Item Generation

As testing increases, in general, and computer-based tests such as MST see wider use, there is a need for more and more test items. This has led to increased interest in the use of automatic item generation (Drasgow, Luecht, & Bennett, 2006). Automatic item generation can take on several different forms. The first, and most straightforward, is identifying an item then identifying features of the items that can be easily manipulated. For example, the numbers in a mathematics problem solving item can be changed to create a different problem. If the problem is intended for elementary school children, then the replacement set of numbers may only be positive integers less than 10, while two-digit integers could be used to generate problems for middle-school students.

This simple replacement of values is the most basic form of automatic item generation. Bejar (1993) discusses the other extreme where item models are devised so that the resulting items would not be easily identifiable as resulting from the same parent item. The goal in this case would be to improve upon test security by testing similar content with items that appear wildly different. This would make it more difficult for examinees to share problems, because the next examinee may see a similar problem and not know it and, therefore, not be able to use the illicit information (Bejar, 1993).

Just as the items generated automatically can vary significantly so, too, can their psychometric properties. Sometimes the goal is to create items that are almost identical in difficulty, while other times, the goal is to provide the item pool with many more items with a wide range of item difficulty. Depending on the potential use of these items, the

psychometric properties of the items may need to be known before they are actually administered. For that reason, it is necessary to make predictions about the item difficulties. There have been two basic approaches to this work. First, cognitive theories are used to identify features of items that are believed to be related to the elusive cognitive complexity (Embretson, 1993) of the item then knowledge of those features can be used to predict item difficulties. Second, some have taken approaches at modeling the difficulty based on the features of the items (Glas & van der Linden, 2001; Sinharay & Johnson, 2008).

There have been attempts to create a difficulty model for a wide range of item types. These attempts have been moderately successful, considering how varied the problems were (Embretson & Daniel, 2008). The narrower the type of problem studied the better the ability to predict the psychometric properties of the automatically generated items.

1.5 Statement of the Problem

For adaptive testing procedures to route examinees through the test, the psychometric properties of each item must be known. However, if testing programs wish to use automatically generated items during test administration there is no opportunity to calibrate the items to determine each item's psychometric properties, so the item properties would need to be predicted. Even though item clones are generated so that they will inherit the psychometric properties of the parent item, these items do not, necessarily, behave exactly as the parent item. Depending on the parent item and the changes made to it, the properties of the cloned item could be very similar or quite different from the parent item. The accuracy of the examinees' scores or proficiency estimate will likely vary depending upon the difference between how items actually function and the predicted psychometric

properties used to score the examinee. The accuracy of examinees' proficiency estimates most likely varies under different MST scenarios.

1.6 Purpose of the Study

The purpose of this study was to determine how accurately examinees' abilities can be estimated when automatic item generation is used in a multistage test. This study examined the effect of using predicted psychometric properties for automatically generated items on the accuracy of examinees' proficiency estimates. For testing programs where assigning a score to each examinee is of utmost concern and the score has important consequences for the individual, the accuracy with which their true proficiency can be recovered is paramount. The study varied several MST conditions including the number of testing stages and the number of items per stage, as well as conditions related to automatically generated items such as percentage of items automatically generated and the variability of the cloned items' predicted properties with their actual properties.

1.7 Significance of the Problem

Whether for reasons of test security, on-demand testing, or simply more testing in general, testing programs, both linear and MST, require more and more items. Item development is both costly and time-consuming. Automatic item generation has the potential to offer significant benefits to these testing programs. In traditional, linear testing, an automatically generated item would be calibrated along with the other items after test administration. However, adaptive tests, including MSTs, use the psychometric properties of items to route examinees through the test. If items are automatically generated and used before calibration, then the psychometric properties of these items must be predicted. The

accuracy of these predictions will affect how accurately examinees' abilities can be estimated. This study quantifies the effect on the examinees' ability estimates when automatically generated items and their predicted psychometric properties are used in MSTs. Test developers can use the results of this study to decide to what extent they can incorporate automatic item generation in their own MSTs.

CHAPTER 2

REVIEW OF LITERATURE

2.1 Introduction

There are many reasons for the shift in popularity from computer adaptive testing to multistage adaptive testing in both education and credentialing. This shift is coupled with an increased need for test items. Automatic item generation can address this need and is a natural fit with computer-based tests as it uses the computer to generate items. Ultimately, the aim of this review of the literature is to consider how automatic item generation can be incorporated in a multistage test. First, we will follow the development of adaptive testing from its informal, oral roots to the computer-based tests of today and will consider the differences in the designs of an item-level adaptive test and a multistage adaptive test. The benefits of using automatic item generation will be discussed, as well as the arguments that automatic item generation and the careful development of item models can support construct validity arguments. Finally, the research into predicting the item difficulties and other psychometric properties of automatically generated items will be considered.

2.2 Adaptive Testing

During an oral examination, an examiner who selects questions based on the respondent's previous answers is administering an adaptive test. This can be as simple as the examiner administering an easy item because the examinee has incorrectly answered a more difficult question (van der Linden, 2008). Administering as efficient a test as possible would be the natural desire of any examiner. A classroom teacher attempting to identify a student's weaknesses would use an informal, adaptive test by asking the student questions that would

hone in on the student's misconception or difficulty. To administer such an exam, the examiner must have some knowledge of the difficulty of the test items, even if this knowledge is informal. van der Linden (2008) contends that because an adaptive test was so natural that the first intelligence test by Binet was an adaptive test. The test itself was completely standardized, and explicitly describes how to select each item based on the examinee's previous responses (DuBois, 1970). Similarly, in 1977 Lord outlined a "broad-range tailored test of verbal ability" that would be suitable for examinees from 5th-grade to graduate school. He noted the psychological benefit of matching the items to the examinee's ability level to avoid frustration or boredom.

2.2.1 Beginnings of Adaptive Testing

Even though the idea of an adaptive test is intuitive and natural, the development of large-scale exams in the United States in the 1940s focused on fixed-length, linear exams (van der Linden, 2008). To administer an adaptive test, the selection of the next item is based on knowing the previous item's level of difficulty as well as the difficulty of the remaining items in order to make an informed selection (van der Linden, 2008). While this could be done informally, as described earlier, to be most efficient this process would need to be automated. Additionally, the issue of scoring is a concern because two examines could be presented with completely different sets of questions in an adaptive test. The determination of an examinee's proficiency is related to total test score and thus an examinee's performance would be directly influenced by the specific exam items selected (Wainer, 1990).Two examinees, taking two different tests, who correctly respond to 17 out of 20 items may have quite different abilities if the items comprising the exams vary in difficulty. In contrast, item response theory uses the difficulty level of each item in

determining the proficiency level of each examinee. The probability distribution of responses to an item is a function of both the proficiency of the examinee and the psychometric properties of the item. An important contribution of item response theory (IRT) is that two examinees can see a different set of items, but these examinees can be placed on the same proficiency scale (Lord & Novick, 1968). This ranking of examinees, "*even if they had not been presented any items in common*" (Wainer, 1990) allows the possibility of a truly adaptive test. Because of IRT, both more- and less-able candidates can be presented with items that are suited to their proficiency level, with the hope that this would relieve some frustration that both groups may feel during testing when presented with items that are too easy or too hard for an examinee's level of proficiency (Lord, 1977). To address the issue of administering a test appropriate for an individual examinee, Lord (1980, chapter 8) described a flexilevel test that could be used as a paper-based item-level adaptive test. The items would be arranged, roughly, in order of difficulty, so that the examinee would begin in the middle and if the first response is correct, the examinee would move to the next most difficult item, if the response was incorrect, the examinee would move to the previous item in the list. Lord described how an examinee would follow these rules using a color-coded system with explicit instructions on how to select the next item. Of course, Lord anticipated that exams would eventually be administered and scored by a computer and thus described tailored testing. The necessarily large pool of items would be calibrated using IRT and the next item to be administered would be selected based on which item contributed the most information based on the item information function (Lord, 1980, chapter 10). Item selection and other technical details of adaptive testing will be discussed later.

2.2.2 Procedures of Adaptive Testing

While it may be intuitive to ask students questions in an adaptive manner, formalizing the procedure is more complicated than simply selecting an item the examiner thinks is an easier or harder item, as needed. There are three stages of adaptive testing: first, an initial estimate of the examinee's proficiency is determined which may be used to determine the difficulty of the starting item; second, the examinee's responses are used to update the proficiency estimate which is then used to select the next item; then the final proficiency estimate is determined (van der Linden & Pashley, 2010). Thissen and Mislevy described each of these three stages with a simple question:

> (1) How to START: What is the first item presented to the examinee?
>
> (2) How to CONTINUE: After each response, what is the next item?
>
> (3) How to STOP: When is the test over? (Thissen & Mislevy, 1990, p. 103).

2.2.2.1 How to Start

An adaptive test works by selecting items based on information from the previous items. However, one item always needs to be selected first. There are several ways to select this first item. While we may not know specifics about a particular examinee, we do know the average level of proficiency for the average examinee in a given population (Thissen & Mislevy, 1990). Therefore, a reasonable starting point would be an item with an average level of difficulty, so that all examinees would have the same starting item. To avoid giving every examinee the same initial item, items of approximately equal difficulty would be used as the starting item to reduce the possibility that the answer to the one and only first item becomes well known (Wainer, 1990).

With only a little collateral information about the examinee, a better proficiency estimate could be obtained. An examinee's grades in a particular subject, teacher recommendations, or even information such as an examinee's age, race, or sex could be used to inform an initial proficiency estimate (Thissen & Mislevy, 1990). A better proficiency estimate simply means that the stopping point could be reached sooner, and thus the process would be more efficient. However, Thissen and Mislevy point out that there are several issues of fairness involved with these approaches. They provide the example of having two starting estimates, rather than one. The examinees would be divided into two groups, those in the lower group would start with the proficiency estimate equal to the mean proficiency of the lower group, the higher group would start with the mean of their group. With each successive item, an individual examinee's proficiency estimate would be less and less dependent on the original proficiency estimate. However, for two examinees with the same response patterns, the examinee from the lower group would always have the lower proficiency estimate. That the proficiency estimates of the stronger members of the weak group would be under-predicted is inherently unfair. Use of demographic information would result in the same group membership issues. Thissen and Mislevy describe the issue as "efficiency versus objectivity" (Thissen & Mislevy, 1990, p. 110).

Secondly, the effect of the initial estimate on the final estimate needs to be taken into consideration. Would examinees with a higher initial proficiency estimate be at an advantage? Thissen and Mislevy (1990) suggested that different starting points could be used, but that a procedure which "forgets" the starting proficiency estimate could be employed. In this way, the differential starting estimates would aid the adaptive procedure in more efficiently selecting items appropriate for the examinee.

Once the first item is selected and responded to, the first ability estimate based on the examinee's response must be made. Ability estimation in adaptive testing poses an issue because, unlike in linear testing, the full complement of the examinee's responses is not available for the usual maximum likelihood estimation (MLE). The main concern is that MLE cannot provide an estimate when the responses are all correct or all incorrect, which would not be that unusual after only a few items have been administered (van der Linden & Pashley, 2010). van der Linden and Pashley propose possible solutions to this problem, which they themselves find unsatisfactory: (1) proficiency estimates can be set a given high or low value if a series of correct or incorrect responses, respectively, are provided, (2) refrain from making proficiency estimates until more items have been administered, and (3) a Bayesian solution or use of collateral information, as described above, could be employed. The authors' concerns are that the first two solutions require arbitrary numbers of items and while the Bayesian and collateral information solutions could be very useful, a mistaken choice of prior could get the algorithm off on the wrong foot and down a longer path than necessary. The collateral information, which is equivalent to using empirical priors, could lead to cultural bias. Fortunately, van der Linden and Pashley claim that if there are more than 20-30 items in a test, the issue of the initial proficiency estimate is moot because the estimator will have enough items in which to get back on track (2010).

2.2.2.2 How to Continue

After each response the proficiency estimate is updated. The next item to be presented is chosen as the optimal item based on the updated proficiency estimate. The item whose difficulty is closest to the current proficiency estimate is, theoretically, the optimal item (van der Linden, 2008; van der Linden & Pashley, 2010). In practice, Fisher's statistical

information function is often used to determine which item, of all available items, would provide the most information. The notion of most information, in a statistical sense, is related to attaining a small standard error associated with the proficiency estimate. One danger in selecting the optimal item is because the earliest proficiency estimates during the exam can be inaccurate, as these estimates are based on so few items, then less than optimal items may be selected. In a sense, this, perhaps, highly discriminating item, was wasted to make an early in the exam estimate that may not be that accurate anyway (Thissen & Mislevy, 1990). If the next item is selected as optimal for the current estimate, which turns out to be far from the true proficiency level, then this "optimal" estimate may be offering very little information with respect to the true proficiency level (van der Linden and Pashley, 2010). For this reason, some item selection procedures take into account the error associated with both the examinee's proficiency estimate and the estimates of the item parameters. Another consideration is that the most informative item remaining in the pool may cover the same content as several items already selected, so it may be necessary to select the most informative item from a set restricted by test content (Wainer, 1990).

2.2.2.3 How to Stop

Theoretically, the item selection step is repeated until the standard error of the proficiency estimate is below some pre-determined level. In practice, however, testing programs often set a time limit and a limit on the number of items in a given examination (Thissen & Mislevy, 1990). It is quite possible, then, that some examinees will reach the end of the test due to either time or item limits without reaching the confidence level on their proficiency estimate. Likewise, it is possible, although not as problematic, that an examinee's proficiency level can be determined with very few items, but more need to be given to the

examinee to meet the minimum requirements. In practice, a combination of target precision, a maximum number of items, and time constraints are used to define the stopping point.

2.2.3 Disadvantages of CAT

There are many benefits to item-level adaptive testing; primarily, the efficiency at which an examinee's proficiency can be obtained, this, in turn, can lead to shorter testing time, a benefit for the examinee and the testing program. In the case of large-scale testing programs, CATs are administered on a computer, so examinees can receive their score reports almost instantaneously. However, there are some disadvantages specifically related to the item-level nature of the adaptive testing.

2.2.3.1 Item Exposure.

One reason the issue of item exposure is more of a concern with CAT than with a fixed-length, linear test is a direct result of some of the advantages of CATs. For example, because CATs are administered on a computer and can be administered on-demand, if the testing program allows, then it is possible that an examinee can take an examination and discuss the exam with next week's examinee. If the item pool is large enough, it is unlikely that the two examinees will see the exact same items, or even see one item in common. The early efforts to control item exposure were attempts to control the percentage of examinees that would see a particular item (Stocking & Lewis, 1998). Even though these constraints were satisfied, there was still a problem with item exposure. Item use is not uniform; some items are presented much more frequently than others (Wainer, 2002). Consider a particularly difficult item that would be administered only to those examinees who have correctly responded to many items. A small percentage of the overall examinees would

qualify to see such an item, so if this item provides particularly useful information, with respect to Fisher's information then based on the algorithms to select the optimal item, this particular item might be selected for *all* of these stronger examinees. This is particularly true if the item pool is lacking in well-behaved difficult items. It is reasonable to imagine that a more able examinee would discuss the exam with other strong examinees for whom there would be a very high likelihood of seeing the same item. To address this issue of item exposure, Stocking and Lewis (1998) recommended item exposure controls conditioned upon the examinee's proficiency. However, this still requires a large item pool, with sufficiently many items across the range of item difficulties.

2.2.3.2 Inability to Evaluate Final Exam Form

The items on a fixed-length, linear test can be thoroughly vetted with respect to item content and how these items work together as a single test, for example, guaranteeing that one item does not provide a clue to the answer of another item on the test and that the content is adequately represented by the selected test items (Luecht & Nungester, 1998). While it is possible to build these features into selection constraints, there are some issues that are more quickly and easily identified by an expert eye (Hendrickson, 2007). Even if there are multiple forms for a given administration, all of these forms can be thoroughly inspected so that the testing program is assured that all examinees will be presented with as similar exams as possible with respect to difficulty, timing, and content. However, in the case of adaptive testing, the test form that each examinee takes will not be constructed until the examinee is sitting at a computer terminal taking the exam.

2.2.3.3 Time Differential in Completing Similar Sets of Items

For two examinees that differ considerably in proficiency, the difficulty of the selected items should also vary considerably, as the nature of an adaptive test would dictate. Even if a strong examinee knows how to solve a difficult problem, these difficult problems often take longer to complete than an easier problem with the same content solved by a weaker examinee (Bridgeman & Cline, 2004). In analyzing the GRE-Analytical, Bridgeman and Cline (2004) found that there was a significant disparity in time to complete, even when the items were of the same content and same difficulty level. Examinees who, by chance, received a series of items whose average time to answer was longer than average, were at a disadvantage. Stronger candidates, for whom speededness may not have been an issue on a fixed-length, linear exam, may find that they need to guess on the last several items of a CAT because they were given an exam that was not comparable with respect to time to complete as the exam given to less-able examinees. These issues related to timing are in contrast to what Green saw as "the most intriguing psychometric advantage" of an adaptive test: that examinees could work at their own pace, leading to an "ideal power test" (Green, 1983, p. 70). For logistical and technical reasons however, many testing programs do have a time limit on each section of their tests, thereby eliminating the potential advantage that Green envisioned. Bridgeman and Cline (2004) suggest making a CAT less speeded to account for the inability to create two sets of items, of even comparable difficulty, that take the same amount of time to complete. They stress that this does not mean removing all time restrictions, but making the test somewhat less speeded.

2.2.3.4 Test Anxiety

The issues related to test anxiety with respect to CATs are not always straightforward. Intuitively, because examinees are not allowed to skip an item or go back to an earlier item, it can be imagined that this would introduce a potentially anxiety producing element not present in a paper-based test (Mead, 2006; Patsula, 1999; Vispoel, Rocklin, & Wang, 1994). On the other hand, arguments can be made where CATs would be associated with a decrease in test anxiety. As Powers (2001) points out, for weaker examinees in a traditional fixed-length, linear test setting, they were surrounded by others who were most likely going to outperform them on the examination. In this same gymnasium there are also the proctors marching up and down the aisles which can contribute to an examinee's anxiety or nerves. Powers makes the case that when taking a CAT, not only are the potentially intimidating fellow examinees reduced, but the test will be at a more appropriate level for the weaker examinees than that paper-based test. The items on a paper-based test were most likely arranged from easy to difficult so that when the weaker examinees were at the end of the exam, they would be faced with items that were far too difficult for their proficiency level. However, because the CAT selects items targeted to the examinee's proficiency level, the weaker examinees would have a better chance of correctly responding to items at the end of a CAT than the final items on a paper-based test, which could be encouraging to the weaker students (Bridgeman, 1998). Conversely, stronger examinees that are used to correctly responding to most items on paper-based tests may be surprised when the final items of the CAT are quite difficult. The issue of test anxiety is not clear-cut; features of CATs may contribute both to increasing or decreasing test anxiety for different test takers.

2.3 Multistage Testing

The different types of computer-based tests can be considered on a continuum of individualization (Jodoin, Zenisky, & Hambleton, 2006). The least individualized design is the computer-based version of a linear paper-based test, often referred to as a linear fixed-length test. At the other end of the continuum is the item-level CAT described in the previous section. The multistage test (MST) is in between as it combines features of both a fixed-length test and the adaptive nature of a CAT. Briefly, an MST adapts to the examinee after a set of items has been completed, rather than after each item, as in an item-level CAT (Luecht & Nungester, 1998; Hendrickson, 2007). These sets of items can be constructed before the administration of the exam, thus allowing for content review (Luecht & Nungester, 1998; Breithaupt, Ariel, & Veldkamp, 2005), for example. The current popularity of MSTs can be attributed, in part, to some of the disadvantages of CATs, such as: test developers not being able to review the test as a whole (Luecht & Nungester, 1998), examinees not being able to review earlier items (Mead, 2006; Patsula, 1999), and the issues related to differences in time to complete ostensibly comparable sets of items (Bridgeman & Cline, 2004).

2.3.1 Early Examples of Multistage Tests

Before examining the current state of multistage testing, we will consider some early examples of MSTs. Cronbach and Gleser (1956) described a two-stage sequential sampling plan for personnel selection (1956, chapter 6) that addressed issues such as the cost of testing, which was frequently ignored in test theory. If resources were no object, Cronbach and Gleser claim that administering a complete set of tests to all candidates would be the most beneficial optional. However, this rarely being the case, Cronbach and Gleser

presented a testing scenario in which individuals would only be given the second stage of testing if their performance on the first stage was adequate. If, for an individual, it was determined that their performance on the first stage precluded their eventual selection, then there was no reason to administer the second stage to the individual. This testing requires a thoughtful design of the first stage to be able to make a determination to eliminate someone from the selection procedure with only that information; this procedure is similar to pre-screening of applicants, which had been done for years. Cronbach and Gleser wished to formalize the procedure and evaluate its improvements in efficiency and reliability.

In their sequential sampling, Cronbach and Gleser only made categorical distinctions, such as selection or rejection. Lord (1971, 1980) was more interested in measurement, than classification, when he wrote of a two-stage testing design in which the first stage was a routing test used to select the second-stage test most appropriate for the examinee's ability level. The most important benefit of this testing design would be for examinees at the extremes of the ability range. Because linear tests are designed to address the middle of the ability range, or the average candidate, the two-stage test would provide a more appropriate set of items for the examinees at the extremes of the ability range. Lord, provides the features of the two-stage test that need to be considered during test development:

(1) The total number of items given to a single examinee (n).

(2) The number of alternative second-stage tests available for use.

(3) The number of alternative responses per item.

(4) The number of items in the routing test (n_1).

(5) The difficulty level of the routing test.

(6) The method of scoring the routing test.

(7) The cutting points for deciding which second-stage test an examinee will take.

(8) The difficulty levels of the second-stage tests.

(9) The method of scoring the entire two-stage procedure (Lord, 1971, p. 288 & Lord, 1980, p. 129).

Even though three decades have passed since this list was written for a paper-based two-stage test, Zenisky, Hambleton, and Luecht (2010) wrote that decisions about these features are still critical when constructing a computerized multi-stage test. Because of developments in multi-stage testing, there are several other aspects that need to be considered: "the number of stages, the ability distribution of the examinee population, the extent of target information overlap for modules within stages, whether random module selection (at an appropriate difficulty level) or panel-based administration is used, whether content balancing is done at the module or total test level, the choice of method for automated test assembly, the size and quality of the item bank, how test information is distributed across the stages, the placement of cut-scores for pass-fail decisions, the issue of item review, and item-exposure levels" (Zenisky, at al., 2010, p. 357).

2.3.2 Overview of MST Design

An MST is an adaptive test with some of the benefits of a paper test. First, the element of the test that allows for adaption is not each item, but a set of items. Each set of items, or a module (Luecht & Nungester, 1998), comprises a stage of the MST. In the literature there are many terms used to describe the same features of an MST, for example, the modules are sometimes referred to as testlets (Wainer & Kiely, 1987). The diagram of an MST in Figure 2.1 displays and labels the main features of the test.

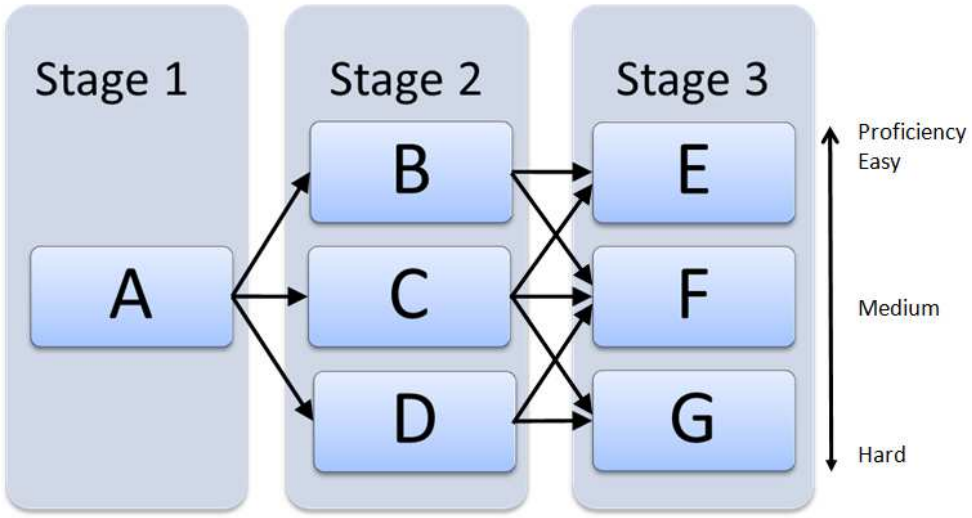


Figure 2.1 Example of 3-Stage MST with 7 Modules A-G

The diagram in Figure 2.1 represents a sample three-stage MST. Module A in Stage 1 would be used for routing all examinees. Based on an examinee's performance in Stage 1 the examinee would be routed to Module B, C, or D, each module would target different proficiency levels (Luecht & Nungester, 1998). After Stage 2, examinees are again routed to a module in Stage 3. In the sample MST shown in Figure 2.1 only adjacent moves are allowed; for example, there is no arrow from Module B to Module G. The decision about possible paths and whether extreme or only adjacent routings are allowed must be made by the test developers.

Each module can be considered as a linear test form in and of itself. Each module can be assembled to meet content specifications and can be inspected before the exam is administered, a significant advantage over the item-level adaptive test. The stage in an MST is the "administrative division" (Zenisky et al., 2010, p. 355) of the exam. After each stage, the examinee's entire performance to that point is used to obtain a proficiency estimate and that estimate is then used to select the most appropriate module in the next stage. In an

attempt to reduce item exposure multiple instances, called panels, of each module are constructed. These panels are designed to be interchangeable so that all examinees who are administered the medium-difficulty module in stage 2, for example, will not necessarily see the same set of items.

An MST often begins with an initial routing module that may contain items across a wide range of difficulty levels or concentrated around the average proficiency level of the expected examinees (Hendrickson, 2007). Based on their performance on the initial module, examinees are routed to modules in subsequent stages that will yield the most precise estimates of their proficiencies. In multistage testing, each subsequent stage has a narrower range of item difficulties. Hendrickson (2007) describes the possible number of stages as ranging from 2 to the number of test items, which would be an item-level CAT. More stages and more modules per stage offer more flexibility, but Luecht and Nungester (1998) found that measurement precision did not improve greatly with increased complexity of the test structure.

Luecht, Brumfield, and Breithaupt (2006) describe the three stages of scoring in an MST. First, the individual items must be scored, next all modules taken by an examinee up to the current point must be scored to select the appropriate module in the next stage, then a final score for the examinee is generated and any pass-fail decisions may be made. This last stage of scoring does not need to happen at the test administration site. If there are technical or security concerns about the amount or type of data required at the administration site, then the final score assignment can occur off-site, while only the information needed to route examinees would be embedded into the program at the site (Luecht, et al., 2006). Number-correct or IRT scoring may be used in MSTs. If using IRT scoring then the IRT

model must be chosen, such as the three-parameter logistic model, the nominal model, or graded response model, for example (Hendrickson, 2007).

2.3.3 Construction of a Multistage Test

2.3.3.1 Assembly of Modules

The purpose of the test, population to be tested, and decisions to be made based on the test, must be considered when deciding upon the specific features of the MST. These considerations will guide the test developers when determining the length of the test, the degree of difficulty for each stage, the stopping rule, and other decisions (Hendrickson, 2007). Development of an MST needs to ensure that any legitimate combination of modules results in a complete test that meets specific statistical and content specifications. Luecht and Nungester described two different MST design strategies: "bottom-up" or "top-down." The bottom-up approach requires module-level statistical and content specifications then multiple modules can be assembled meeting those specifications. The different modules can then be distributed across different panels. The top-down approach selects among modules so that the overall test meets test-level statistical and content specifications. It is more difficult in the top-down approach to exchange modules across panels, because modules across panels will not, necessarily, meet the same specifications, but are designed to work with the other modules within one panel (Luecht & Nungester, 1998).

The three general steps for assembling panels from modules were described by Luecht and Nungester (1998): (1) generate the statistical targets for testlets across different stages, (2) allocate the content specifications across stages, and (3) one panel at-a-time assemble the modules that meet the constraints of steps 1 and 2. For the first step, Luecht and Nungester recommend using IRT test information functions as targets for each module

or a combination of modules. The target test information function states the amount of measurement precision desired across the proficiency scale. The conditional error variance of the proficiency estimate is the reciprocal of the test information function (Luecht & Nungester, 1998). Based on how much error the test developers are willing to accept at specific ranges across the proficiency scale will dictate the shape of the test information function. In the top down approach, a target information function can be developed for one module or a combination of modules which represent a particular path through the exam. Test developers may want to consider the most probable paths for the majority of examinees and develop a test information function for each likely path (Luecht & Nungester, 1998). Based on the MST shown in Figure 2.1, examinees of average proficiency would most likely follow the A-C-F path, weaker examinees could follow A-B-E or A-C-E, and stronger examinees would most likely see modules A-D-G. The test developer needs to consider whether they will only allow adjacent moves or if examinees can make more extreme moves. This method of establishing a global test information function for each probable path is consistent with top-down strategy. Conversely, the bottom-up strategy consists of developing target information functions for each testlet. Under both approaches, test content specifications and such features as item type need to be included in the specifications when assembling the modules and, ultimately, the panels. Each target information function reaches a maximum value at the range on the proficiency scale where the greatest measurement precision is desired (Luecht & Nungester, 1998).

Hendrickson (2007) describes the choice of statistical targets for the modules as "one of the most important decisions in designing the [MST]" (p. 49). The average level of difficulty and the range of difficulty for each module must be determined. Maximizing the test's information is the goal, while maintaining the desired shape of the test information

function. The statistical properties of the routing, or first-stage, module were highly related to the measurement precision of the test as a whole. In a study of two-stage tests, Kim and Plake (1993) found that if the routing module had a wide range of item difficulties there was ultimately better measurement precision at the ends of the ability distribution, while a routing module with item difficulties focused around a particular level had better measurement precision in the middle of the ability distribution, but that the extent of the differences varied based on the number of modules in the second stage.

2.3.3.2 Scoring and Routing

As noted earlier, it was the development of IRT and the ability to assign difficulty levels to particular items that allowed adaptive testing to take place on a large scale. Therefore, it seems logical to assume that IRT scoring would be most appropriate for routing purposes and generating a final estimate of proficiency for the examinee. In IRT scoring, a maximum likelihood procedure uses the response pattern and item statistics to determine the most proficiency level that maximizes an examinee's responses. However, Luecht and Nungester (1998) demonstrated that number correct scoring was sufficient, in most cases, for assigning an examinee to the appropriate module in the next stage.

The scoring that occurs during the administration of the MST is used for routing. Whether using number correct or IRT scoring, the test developers must decide where the cut points will be for routing examinees to the appropriate module. Two methods for determining these cut points are described in the next section. For now, let's assume that a cut point, or decision point, has been selected, θ_d, along the proficiency scale. Again, using the sample test shown in Figure 2.1, this cut point will be used to route examinees to Module B or C based on their Stage 1 performance. The examinee's pattern of item responses for

Module A will be used to find the examinee's estimated θ on the proficiency scale using the maximum likelihood procedure described above. If the examinee's estimated θ is less than θ_d then the examinee would be routed to Module B, otherwise the examinee is routed to Module C (Luecht & Nungester, 1998).

In the case of number correct scoring, the cut point on the proficiency scale, θ_d, is again used. The expected number of items correct for an examinee at the chosen proficiency level, θ_d, can be calculated using the IRT item parameters for the items in Module A, the chosen proficiency level, θ_d, and the selected item response function, such as three-parameter logistic function. This value, the predicted number of items correct, can then be used for routing to the next module. If an examinee responded correctly to fewer items than predicted for someone with a proficiency level of θ_d, the examinee would be routed to Module B, otherwise Module C. The number correct scoring ignores the specific response pattern of each examinee, in other words two examinees who both correctly answered 13 out of 20 items in Module A would be routed identically, regardless of which items each examinee correctly answered (Luecht, et al., 2006).

2.3.3.3 Determining cut points for Routing

Before routing can occur, the cut points which dictate the routing must be established. To compute these cut points in terms of number correct or even using a strictly IRT procedure, the test developers need to determine where those cut points are along the proficiency scale. Luecht et al. (2006) described two procedures: the approximate maximum information method and the defined population intervals method. The approximate maximum information uses the cumulative test information function based on the testlets so

far and the current testlet to determine which testlet in the next stage would offer the maximum information for a given examinee based on the examinee's current proficiency estimate. For example, one of the cut points to be used after Stage 1 would be at the intersection of the target information functions for the items along path A-B and the items along path A-C, the easy and medium paths through Stage 2. The two intersecting target information functions are shown in Figure 2.2, along with the intersection point and, therefore, the lower cut point. Examinees whose estimated proficiency falls below the cut point would be routed to Testlet B in Stage 2, because the A-B path would provide more information for those examinees than the A-C path.

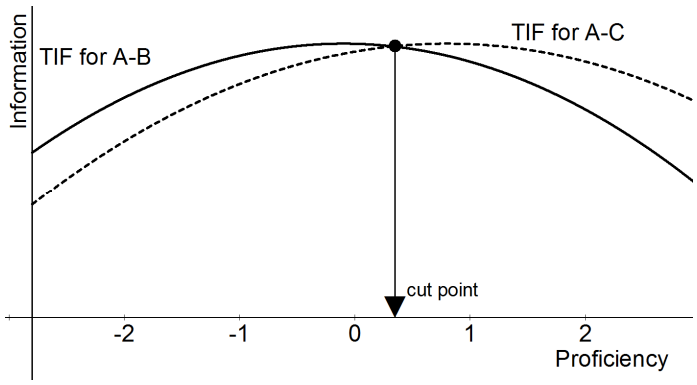


Figure 2.2 Intersection of Test Information Curves for Paths A-B and A-C

Similarly, the intersection of the test information curves for the items along paths A-C and A-D would provide the other cut point. After completing Stage 1, examinees whose estimated proficiency level is below the 2nd cut point would be routed to Testlet C and those with an estimated proficiency level is greater than the cut point would be routed to Testlet D. This process is similar to an item-level adaptive test selecting the next item to maximize information.

The second option described by Luecht et al. (2006) is the defined population intervals method, which specifies the proportion of examinees that are expected to follow

specific paths. This would be a policy decision and requires an assumption of the distribution of proficiency estimates. Using the defined proportions, percentiles can be computed, and find the corresponding proficiency level based on the proficiency distribution. As in approximate maximum information, the expected number correct would be calculated for each proficiency level, as described above.

2.3.4 Advantages of MSTs

While there are many advantages to MSTs, there are several disadvantages, when compared with CATs. Due to the change from item-level adaptive to adaptive by stages, there is a decrease in efficiency and accuracy. The greatest advantage of a CAT is the ability to present the examinee with the item that is most likely, mathematically, to provide a more accurate estimate of the examinee's proficiency, in other words more items are needed with MST to obtain the same measurement precision as a CAT. However, the differences in measurement precision between CAT and MST are not great; MST has improved measurement efficiency when compared with a linear test of the same length (Hendrickson, 2007). The number of stages and number of difficulty levels within each stage affects the precision of the estimates. For many testing programs, the advantages of an MST, such as the ability for an examinee to move or skip around and review items within a module, the ability to prescreen a module for test content, and the increased control over selecting items that take the same amount of time to answer outweigh the slight loss of measurement efficiency when compared with a CAT.

2.3.4.1 Evaluating Test Forms

Unlike item-level CATs, MST affords the opportunity to address quality assurance before test administration (Luecht & Nungester, 1998). The modules and combinations of modules which will comprise the complete tests may be reviewed by experts for problems that may not be detected during an automated assembly process. Even though CATs use explicit criteria when selecting an item, which may include constraints related to statistical characteristics of the items, content-related constraints, there are some features that may not be obvious until a human examines the set of test items as a whole. Building all possible constraints related to item ordering and context effects could not be reasonably implemented (Hendrickson, 2007). It may take an expert eye to look at the set of items making up the testlet to realize that while a series of items may cover different content the scenarios may be similar as to cause some undue confusion.

2.3.4.2 Technical and Statistical Advantages

Local independence can be better controlled among sets of modules in multistage testing than in an item-level adaptive test (Thissen, Steinberg, & Mooney, 1989). A module that is based on common content or a common stimulus can be scored using a polytomous model which does not have the requirement of independence (Hendrickson, 2007).

Statistical analyses, such as detecting differential item functioning, are easier to conduct under MST because of the structure of the data and because the examinees who respond to each item are more heterogeneous than in CAT; the missing data is in blocks rather than the sparse data, or non-missing data scattered throughout, as it would be in CAT (Mead, 2006).

2.3.4.3 Examinees' Ability to Review Items Within a Module

Examinees most common complaint about item-level CATs is the inability to skip an item then return to it later and the inability to go back and review items (Vispoel et al., 1994). In MSTs, examinees have the ability to move around within each module. Unlike the item-level CATs, because routing does not occur after each item, an examinee can continue to adjust his or her responses until the module is completed. After the module is completed the examinee's responses will be used to select the module for the next stage of the exam. For many examinees the feature afforded by MSTs to review items within a module reduces test anxiety when compared to item-level CATs (Mead, 2006; Patsula, 1999).

2.4 Automatic Item Generation

As described earlier, CATs and MSTs, especially when administered frequently, need to maintain a large-item pool (Breithaupt, Ariel, & Veldkamp, 2005, Wainer, 1990). For this reason, automatically generated items are attractive to many testing programs. In adaptive testing, the psychometric properties of the items must be known so that the next item or module can be selected appropriately. The items could be calibrated ahead of time, but this can be very expensive and perhaps not viable if only small samples are available. If items are to be generated during test administration, then reasonable estimates of item difficulty must be available for the algorithm to select the next item or for scoring of examinees to occur. If it is necessary to administer and calibrate all items in order to obtain their difficulty, then there are neither great cost-savings nor improvements in efficiency by automatically generating items. However, if items can be automatically generated and their difficulty levels predicted, then there, then there would be a more efficient system.

Automatically generating items is possible, but predicting their psychometric properties is much more difficult. There have been many attempts to identify features of quantitative items that contribute to item difficulty. The ability to predict item difficulty without the need to pre-test or calibrate items based on the responses of 100s or 1000s of examinees could lead to tremendous savings in both cost and time. Attempts to identify generic item attributes that contribute to item difficulty often used a restrictive set of item types. However, some of these attempts have resulted in models that could predict item difficulty, even if only for a specific item type. In this section we will describe why we would want to model or predict item difficulty and consider the role that these item models play in establishing construct validity. Then we will outline the existing theories for different quantitative item types. Many of the studies in this section use items from the Graduate Record Examination (GRE), most often the Quantitative section.

2.4.1 Benefits of Automatic Item Generation

The ability to automatically generate items could reduce some of the issues of item exposure. Because many items produced by human item writers are discarded before operational use, it would be advantageous to generate items automatically to meet specific criteria and that have specific psychometric properties. Items generated automatically can be clones of each other or variants. Item clones, or isomorphs, are generated with the goal of creating items with psychometric properties identical to the original item. With item variants, on the other hand, the goal is to generate items with a wide range of psychometric properties, specifically with a variation in item difficulty. It has also been a desire of some researchers to develop strategies for automatic item generation that will produce items that appear quite different to the examinees yet measure the same underlying construct. This

could play an important role with respect to test security: even if one examinee describes an item to a future examinee, the item on the future exam may not be recognizable and the prior information would be worthless.

2.4.2 Automatic Item Generation and Construct Validity

One can imagine two almost identical arithmetic items, one involving subtraction with "borrowing" and one without. All things being equal, for students just learning how to subtract, subtraction with carrying will be more difficult. However, the uncertainty arises when a variety of items are combined in one test. The distinction in item difficulties is determined by calibrating the items after administration. This is unsatisfactory from a construct validity point of view, because, intuitively, item difficulties should be explicitly related to the difficulty of the content. In practice, we assume this is the case when we determine item difficulties by calibration after administering the items to sufficiently many examinees. By definition, of course, this is true: Item difficulties are determined by which items are more difficult for the examinees. However, this process does not help test developers identify what makes an item difficult.

Bejar (1993) addressed the concern of some that IRT ignores the underlying construct the test is intended to measure and, instead, only focuses on the modeling of the responses. He suggested that the process of item generation coupled with attempts to model the difficulty of the resulting item, how the item interacts with the construct is addressed. The ability to predict what makes an item difficult is a great contribution to construct validation (Mislevy, Sheehan, & Wingersky, 1993). One perspective is that items can be automatically generated by developing item models such that item features are purposefully tweaked so that specific aspects of the construct are tested and the item difficulties can be

predicted based on this process. Some cognitive psychological research relates the psychometric properties, including item difficulties, of automatically generated items with their parent items, or the items from which they are derived. The concept of item generation would allow more attention to be paid, not only to the content in terms of test specifications, but to the underlying construct itself. If the research in developing item generative methods results in a deeper understanding of what processes an item elicits from the examinee, then there is the potential of developing an exam truly measuring the underlying construct of interest (Embretson, 1999).

Before items can be automatically generated, we must have a better understanding of the nature of the item itself. In Embretson's (1999) research on automatic item generation she referred to her earlier work on validity. Embretson (1983) proposed a distinction between construct validity and nomothetic span. She claims that construct validity refers to the processes and thinking that examinees go through in order to solve a problem, while nomothetic span refers to the relationship between the current instrument and other instruments that purport to measure the underlying construct. While the desire to generate items automatically may have started as practical in nature, to meet increased demand for items or as an attempt to improve test security, the improvement of automatic item generation will serve to further our understanding of and improve the validation of the underlying construct.

2.4.3 Exploring Factors Related to Item Difficulty

In an ideal situation, a computer adaptive exam could manipulate a complex set of features to automatically generate an item tailored to the current needs of the test. Current needs could refer to everything from psychometric needs, in terms of item difficulty and

discrimination, to content representation or even item format. Enright and Sheehan (2002) described a collection of studies that used GRE quantitative items to isolate the factors that contribute to item difficulty. The first study they discuss is an attempt to model cognitive skills and the properties underlying quantitative GRE word problems. After analysis of student work and various solutions to the word problems, Sebrechts, Enright, Bennett, and Martin (1996) coded each problem in order to relate the mathematical features of each problem to its corresponding difficulty. Using multiple regression, they found 3 features related to item difficulty: (1) algebraic manipulation, that is whether variables needed to be manipulated because there was more than one variable in the representation of the problem, (2) item content, such as money or time and distance, and (3) mathematical complexity, referring to levels of nesting, the mathematical structure of the problem, and number of operations. Multiple regression using two sets of item difficulties were used in this study: first, the item difficulties based on GRE examinees taking the set of items in a multiple-choice format and second, the item difficulties based on college students taking the set of items in constructed response format. Despite the difference in item format, the variability of item difficulty explained by the three factors listed above ranged from about 37% to 62%. Additionally, they noted that items with the same difficulty did not, necessarily, share the same characteristics, this provides a note of caution with respect to our expectations of item generation.

Next, Enright and Sheehan detail a study that systematically varies features of items so that the contribution to item difficulty of particular aspects of a word problem can be studied (Enright, Morley, & Sheehan, 2002). They note that in using extant GRE quantitative word problems, Sebrechts et al. (1996) were limited to the variations that were already in the problems. In an experimental situation, the content, complexity, and use of variable, could

be controlled in each problem. The study limited the type of word problem to rate and probability items. The content of the rate problems were either money or distance/time. The content of the probability problems had 2 layers: (1) percent or probability and (2) abstract or real-life scenario. In addition, the use of a variable was varied for the rate problems and the complexity level was varied for both types of problems by increasing the number of mathematical constraints in the problem. The results showed that these features did, indeed, relate to item difficulty, however, the impact of the features varied for the two types of problems. For example, content in the rate problems was important when no variable was used: for two problems requiring the same mathematical process to solve, those formulated in the context of money were easier for the examinees, however when a more algebraic solution was required there was no practical difference due to content. In the probability problems, whether the problem was abstract or contextualized did not affect the difficulty, however items phrased as a probability rather than a proportion problem were more difficult. While this work was important in identifying aspects of the domain of quantitative reasoning, unfortunately, the results are narrow and seem specific to these two types of problems. Even within the context of just two types of GRE quantitative problems, the importance of context was quite different. This study seems to indicate that even if a cognitive model with enough details to adequately predict item difficulty existed, it would be rather large and cumbersome.

A broad approach to identify the features underlying cognitive complexity in problem solving items was attempted by Embreston and Daniel (2008). Raters classified the cognitive complexity of items according to such criteria as the number of words, terms, and operators in the items stem, whether the examinee was required to supply an equation to

solve the problem, the grade levels of the knowledge and the computations procedures required, and the number of computations required to solve the problem.

Embretson and Daniel studied the proposed cognitive complexity variables in terms of predicting item difficulty. Fischer's (1973) linear logistic test model (LLTM) and a regression model were both used to predict item difficulty. They used GRE quantitative items in this study and these items varied considerably in terms of "content, syntax and form, as well as in mathematical requirements" (Embretson & Daniel, 2008, p. 342). Because of this variability, it is even more impressive that the hypothesized cognitive model worked so well, in fact, about half of the variance of item difficulty was attributable to the cognitive model. In a previous study the same authors found that systematically varying some of the cognitive complexity variables were associated with changes in item difficulty. The authors propose that if items were generated by altering these cognitive complexity variables then even more accurate predictions of item difficulty could be achieved. Research in item models has shown that item prediction is improved when items are generated by systematically varying features of the item model (Enright, Morley, and Sheehan, 2002; Bejar et al. 2003).

Daniel and Embretson (2010) continued their work, by selecting items in which it would be feasible to systematically vary two of the cognitive complexity variables. Again, they had the goal of being able to incorporate the elusive "cognitive complexity" into mathematical problem solving items. The first variable is whether the equation necessary to solve a particular problem was provided or whether it needed to be supplied by the examinee. Second, they varied the number of subgoals required to solve each problem. Their results supported the notion of incorporating a cognitive model in attempts to generate items with a targeted item difficulty. However, they did suggest that the assumption that an

equation given in words, rather than symbols, would require more processing from the examinee and would therefore make the item more difficult may not hold true.

One possible issue in the research in predicting item difficulty based on cognitive models is the coding of the items. Initially, the categories in these studies seem objective, but in reality, these could be more subjective. For example, different examinees may approach the same problem very differently and thus use a different number of subgoals or a different set of equations to solve the problem. There is also the issue of the *expert blind spot* as described by Nathan and Petrosino (2003), whereby subject matter experts may be overly influenced by the structure of the subject itself and may not consider or have knowledge of how learners actually progress through a subject when making determinations of item difficulty. This blind spot could lead to some degree of bias when positing cognitive models that are related to item difficulty.

In studying items from the, now retired, analytical reasoning section of the GRE, Newstead, Bradon, Handley, Dennis, and Evans declared that "it is clear that a single model is not possible" (2006, p. 87). Instead, they developed separate models for each type of analytical reasoning item. They suggest that a combination of models may be used to better model item difficulty. Similarly, Bejar (2010) asserts that a unified model to predict psychometric properties of items is unrealistic, especially considering the specialization or detail required to make the model meaningful.

Determining what exactly makes an item difficult is a complex and elusive issue. Using the mathematical structure of items, Arendasy, Sommer, and Ponocny (2005) developed a cognitive based model for quantitative comparisons often seen by elementary students. Prior research has shown that whether "more than" or "less than" is used can affect the item's difficulty even if the phrases have the same mathematical meaning, for

example: "A has 5 more than B" may function differently than "B has 5 less than A." Arendasy et al. (2005) considered the process that children use when reading these types of problems, specifically, the children image how they would solve the problem arithmetically in terms of selecting an arithmetic operation as they read the words. Even though this research is related to children and simple arithmetic comparisons, rather than adults solving GRE quantitative items, the lesson is relevant: only considering the underlying mathematical structure of the problem itself could result in ignoring other features of the item which contribute to its difficulty. Conversely, Bejar (2010) warned that just because some factors are not strong predictors of item statistics, they may still be important design features to include in item models.

2.4.4 Item Models

In order to realize the full benefits of automatic item generation, the psychometric properties of the generated items would need to be predicted rather than calibrated. If different models are required to predict item difficulty in items that appear, at first glance, to be so similar, then it becomes apparent that a unified theory is unrealistic. Embretson and Daniel (2008, 2010) attempted a broad approach at characterizing quantitative word problems. The types of problems they considered were by no means exhaustive in terms of quantitative items. If one could develop such a unified model for all quantitative items then the number of parameters required would be so great as to render the model useless. Instead, we have the concept of item models in which there is a separate model for each item type. A specific testing program could develop item families based on historical data and develop an algorithm for generating isomorphs or variants for each item type.

In a study to investigate items features that could contribute to predicting item difficulty, existing items with linear and simple rational equations were categorized according to several features, including whether the stem and key of the item were verbal, symbolic or strictly numeric (Deane, Graf, Higgins, Futagi, & Lawless, 2006). The intent of this exploratory work was to then develop task models that would incorporate these features. Graf, Peterson, Steffen, and Lawless (2005) described the development of generative item models that could predict psychometric parameters. They concluded that the constraints built into the item models would need to be rather restrictive to achieve the best predictions of the psychometric parameters.

Of course, a testing program could simply use automatic item generation because it needed a large number of items and pre-testing or calibrating the items after test administration was not a difficulty. However, if a program wanted to generate items on the fly to use in an adaptive testing situation, then it would be necessary to predict the psychometric properties of the items. Fischer (1973) and Embretson (1999) both propose IRT-based models that incorporate features of the item design. Embretson extended Fischer's model by including terms to model the item discrimination as well as the item difficulty. If we consider the transition from classical test theory, with its focus on total test score, to IRT, which incorporated the examinees' responses to individual items with varying difficulty, then the movement to a modeling approach that focuses not only on a few statistical characteristics of an item, but on the design features that generated that item, is a natural progression (Embretson, 1999).

2.4.5 Variability of Item Parameters under Item Generation

An alternative to the models based on cognitive theory, is a hierarchical approach which imposes an additional level of variability on the item statistics. Glas and van der Linden (2001, 2003) proposed a hierarchical model in which the item parameters for items generated from a specific parent would come from a multivariate normal distribution. Using data from automatically generated items that were administered as pre-test items to actual GRE examinees, Sinharay and Johnson (2008) used the hierarchical model of Glas and van der Linden, among other models, to evaluate the variability of item parameters within item families. Sinharay and Johnson found that the variability differed across different item families. In fact, there was some quite surprising variation within some item siblings, that is, the collection of items generated from a particular item model.

Several studies have found modest increases in the standard errors of the estimates of examinee proficiency when item difficulties were predicted rather than calibrated. However, these increases were not so large that they could not, in most cases, be compensated for by modest increases in the number of test items (Bejar, et al., 2003; Embretson 1999). There is potential to capitalize on this finding by implementing a computer-based exam that uses some or all automatically generated items and their predicted item statistics. This could be especially helpful in an adaptive setting where the item difficulty is needed to determine the examinee's next item or set of items.

2.4.6 Effect of Item Parameter Uncertainty on Examinee Ability Estimates

While knowledge of the item parameters is needed to implement adaptive testing, item parameters have error in that they are calibrated from a sample (Embretson, 1999) yet they are treated as truth. The appropriateness of the sample is reflected in the standard error

of the item parameter estimates. On the other hand, if item parameters are not calibrated, but predicted, there is the potential for even more error in the ability estimates. In a study using mathematical models and cognitive theory, Embretson found that predicted item parameters only correlated in the .70s and .80s with calibrated item parameters, while the correlation was below .50 when the model was based on item features. Ability estimates contain two sources of error (Embretson, 1999). As ability estimates are based on items, then the error in item parameter estimates, whether calibrated or predicted, is manifested in the ability estimate. Secondly, the error associated with the difference between true ability and estimated ability is still a factor. To study the effect of the uncertainty of item parameters on ability estimates, Embretson conducted a simulation study incorporating different levels of uncertainty in the item parameters. She found that the uncertainty did lead to increased bias and decreased precision in the ability estimates, however she concluded that by increasing test length by only a few items the adverse effects of using estimated item parameters could be overcome (Embretson, 1999).

Mislevy, Sheehan, and Wingersky (1993), in their article with the intriguing title: "How to Equate Tests with Little or No Data," describe how to use collateral information to equate with imperfect or incomplete information. Their suggestions will be useful in considering how to deal with the equating of tests for which we predict, not calibrate, the item statistics. Of course, we need to remind ourselves that the calibrated items statistics that we use every day are only estimates and not truth.

2.5 Summary of Literature

One of the benefits of a multistage adaptive test over an item-level test is the ability to screen the panels of items so that all combinations of panels that could possibly comprise

an exam for a specific examinee would be evaluated. Even with items generated on-the-fly, this is still possible. The specifications of the item models would need to include not only the cognitive item features discussed earlier but other properties such as item content, the format and content of the distractors, use of visuals, etc. This would allow prior inspection of the panels, at least in a general sense, without even seeing the items, because they are yet to be created. However, if the constraints are comprehensive enough it could ensure adequate content coverage.

Given the potential advantages of using automatic item generation with a multistage adaptive test, this study evaluated the efficacy of incorporating automatically generated items and their hypothesized psychometric properties in an MST. The work of Sinharay and Johnson (2008) and Enright, Morley, and Sheehan (2002), among others, documented the variability of item difficulty within item families. Their results were used as the basis for simulating item parameters for automatically generated items within an item family. These simulated item parameters were used to generate examinees' responses, while the item parameters from the parent items were used for selecting the appropriate path through the adaptive test and for scoring. The examinees' proficiency estimates based on this procedure were evaluated for accuracy and bias against the examinees' true proficiency levels. Many conditions of the multistage test design were examined: the percentage of automatically generated test items, the number of panels, and the number of items per panel, to name a few. Test developers could use the results of this study to determine the level of automatic item generation they are willing to incorporate into their MST while maintaining a specified degree of accuracy for the examinees' proficiency levels.

CHAPTER 3

METHODOLOGY

3.1 Introduction

In this section the multistage test (MST) conditions, data, procedures, and evaluation criteria for the simulation study will be addressed. The goal of this study was to determine the accuracy with which examinees' proficiencies can be estimated when automatic item generation is incorporated in an MST. Various features of MSTs were manipulated as well as the percentage of cloned items and the variability of the predicted psychometric properties of the cloned items. To conduct as realistic a simulation study as possible, the most commonly employed MST scenarios were used and the psychometric properties of the simulated items were based on the behavior of actual item clones. Because item response theory (IRT) underlies the MST procedures, before proceeding to the methodology, the IRT model to be used will be discussed.

3.2 Item Response Theory

IRT relates an examinee's performance on a particular item to an underlying trait or proficiency, often referred to as θ. Many IRT models exist, but all contain at least one parameter related to the item and at least one parameter related to the examinee (Hambleton, Swaminathan, & Rogers, 1991). IRT allows examinees and items to be placed along the same proficiency scale, where higher numbers indicate more difficult items and more proficient examinees. IRT's 3-parameter logistic model (3PL) will be used in this study:

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}} \text{ (Hambleton, et al., 1991, p. 17),} \qquad (3.1)$$

where a_i, b_i, and c_i are the parameters for item *i* and θ is the examinee's proficiency. $P_i(\theta)$ is the probability that an examinee with proficiency θ will correctly respond to item *i*. The item parameter of most interest in this study is the *b*-parameter, known as the item difficulty parameter. The *a*-parameter is the discrimination parameter. The *c*-parameter is sometimes known as the pseudo-guessing parameter because it comes into play, for example, in a multiple-choice item where examinees with very low proficiency would have a greater than 0 chance of guessing the correct answer.

3.3 MST Design

Before the details of item properties and scoring are discussed, the outline of the test designs will be described. In this study, 2- and 3-stage tests were used because these are the most popular in practice (Luecht & Nungester, 1998). Both the 2-and 3-stage tests started with a routing test in the first stage. The 2-stage designs are shown in Figure 3.1 where the second stage had either two or three modules.

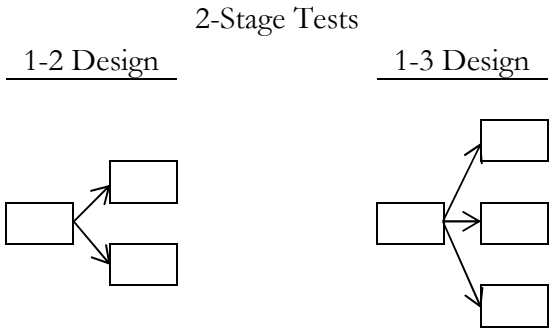


Figure 3.1 Number of Modules in the 2-Stage Design.

The 3-stage 1-3-5 and 1-2-4 configurations of are shown in Figure 3.2.

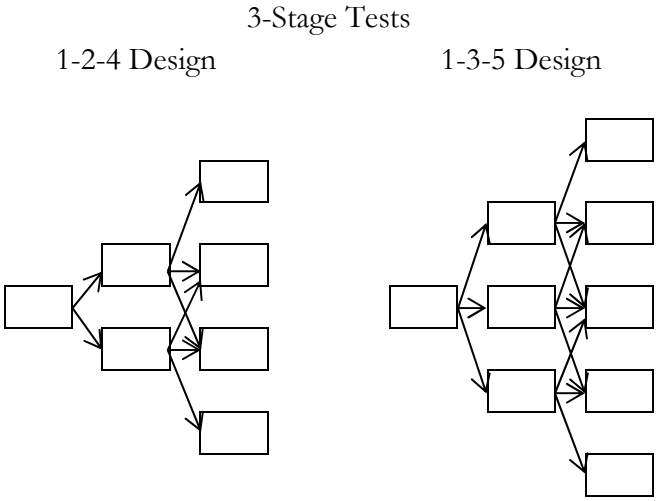


Figure 3.2 Number of Modules in the 3-Stage Design.

There were a total of 36 items in all test designs, 18 items per stage in 2-stage tests and 12 items per stage in 3-stage tests.

3.4 Target Test Information Functions

A target test information function (TIF) was constructed for each module in the MST. The TIF for a given module reflected the desired increased information, and therefore have maximum values, around the cut points for routing for that particular stage. If there are

two cut points after a particular stage, for example, then the TIF for that module should be slightly bimodal. However, the TIFs for the individual modules were constructed so that the overall TIF provides adequate information along the proficiency scale and, ideally, would be relatively flat.

3.5 Item Statistics

Due to the adaptive nature of the MST, estimates of the items' parameters were needed to route an examinee through the test. Conventionally, items with known item statistics are used in MST. The item statistics are known because the items have been previously administered and calibrated, where calibration is the process of estimating the item and examinee parameters. In this study, we simulated tests containing both items that have been previously administered and calibrated and item clones that were generated at the time of administration and would not have been calibrated before use. The item statistics for the clones therefore needed to be predicted. The known item statistics came from an operational large-scale assessment; the descriptive statistics for each module are shown in Table 3.1.

Table 3.1 Descriptive Statistics for Item Parameters in 1-3 MST Design

Mean (Standard Deviation) of Item Parameters Per Module			
Module	*a*	*b*	*c*
Stage 1	1.011 (0.15)	0.042 (1.13)	0.144 (0.08)
Stage 2 Low	1.043 (0.23)	-2.066 (0.60)	0.185 (0.08)
Stage 2 Medium	1.177 (0.27)	0.068 (0.68)	0.148 (0.08)
Stage 3 High	1.090 (0.14)	2.050 (0.530	0.144 (0.12)

Figure 3.3 shows the TIFs for the three possible paths in the two-stage 1-3 MST design. The TIFs for all possible paths in the other three MST designs are in Appendix A.

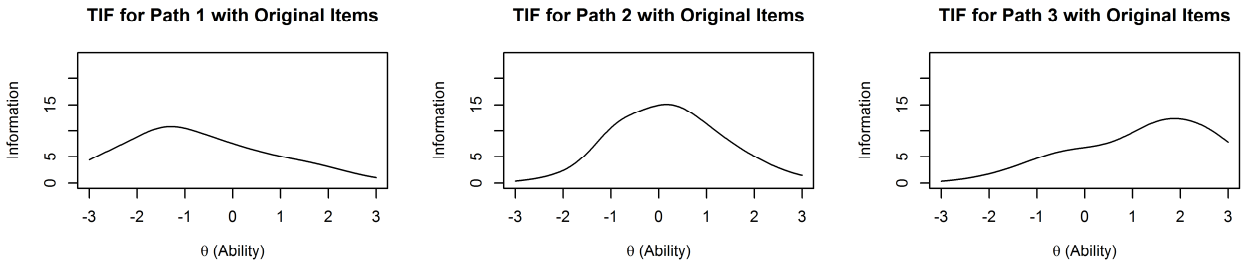


Figure 3.3 TIFs for All 3 Paths in 1-3 MST design

For this simulation study, the item clones required two sets of item statistics. The objective of generating item clones is to create items with the same item statistics as the parent item. The first set of statistics was the clone's predicted item statistics. The predicted item statistics were the same as the parent's item statistics, for all 3 parameters. The predicted statistics were used for the purpose of establishing the cut points used for routing just as the calibrated item statistics were used for the regular non-cloned items.

However, regardless of the efforts to generate operationally identical clones, there is no guarantee that an item clone will function as its parent does: The clone's true item difficulty may differ slightly or considerably from that of the parent item. To account for the potential difference between how the item clone is expected to behave and how it does behave, each item clone will have a second set of item statistics that reflects how the clone actually behaves. In other words, how difficult the examinees find the item clone, regardless of how difficult the item was predicted to be. This second set of statistics differed from the first set of statistics in both the slope and difficulty parameters. This second set of item

statistics were used to simulate examinees' responses to the cloned items as these statistics will represent the items' true difficulties.

The pseudo-guessing parameter for each cloned item was inherited from its parent. The cloned item's difficulty and slope parameters were randomly generated according to previous research on the behavior of cloned items. The results of Sinharay and Johnson's investigation into item clones that were administered in an experimental section of the Graduate Record Examination (GRE) were used (Sinharay & Johnson, 2008). Sinharay and Johnson's observed variances of the difficulty and the log-slope parameters were used to establish the deviations for the cloned items in this simulation, thus modeling the variability of cloned items' slope and difficulty levels when compared with the parent items.

Different parent items produced different degrees of variation in difficulty among their respective sets of item siblings. Some parent items generated sets of item siblings with almost identical item difficulties, while other parent items generated sets of item siblings whose item difficulties were substantially different from the parent and from each other. The difference between a parent item's difficulty and the range of its clones' true difficulties were categorized into 3 groups: (1) small – within 0.2, (2) moderate – greater than 0.2 and less than 0.4, and (3) large – greater than 0.4 and less than 0.6, on the θ proficiency scale, according to the variation observed in the Sinharay and Johnson study.

The variation of the log-slope parameter was relatively consistent over sets of item siblings and was not clearly associated with the variability of the difficulty parameter. For this reason, all item clones were simulated under one condition of variability of slope parameters. To replicate the behavior of the log-slope parameter from the Sinharay and Johnson study a random number from a uniform distribution from -0.33 to 0.33 was added to the logarithm

of a parent item's slope parameter. The exponentiated result of this sum will be the slope parameter of the item clone.

Within each of the 4 test designs, both the percentage of item clones and the magnitude of the clones' variation in item difficulty were varied for a total of 14 conditions, in addition to the baseline condition of no cloned items. In addition to two conditions simulating half and all of the items per stage as cloned items (50 and 100 percent), a condition with a smaller number of cloned items was also tested with the thought that it would introduce only a small amount of variability. Because the two-stage tests and three-stage tests have 18 and 12 items per stage, respectively, a condition of one-third of the items, referred to as 33 percent, was reasonable. Six conditions related to variation in item difficulty were tested: (1) item clones from all 3 variability categories, (2) only clones from the small category, (3) only clones from the moderate category, (4) only clones from the large category, (5) only clones from the large category in the first stage for a 2-stage test and the first and second stages for a 3-stage test followed by clones from the moderate category in the final stage, and (6) only clones of moderate variability in the first stage and no clones in the subsequent stages. The specific conditions with respect to variations in item difficulty and the percentage of item clones are shown in Table 3.2.

Table 3.2 Conditions of Item Clones per Test Design

Percentage of Clones	Variability of Clones
33, 50, 100	Small
50, 100	Moderate
33, 50, 100	Large
50, 100	Small, Moderate, Large (equal numbers)
50, 100	Moderate (final stage), Large (1st and/or 2nd)
50, 100	Moderate (1st stage), no clones in later stages

3.6 Examinee Data

For each replication of an examination 1000 examinees were simulated at 61 locations evenly spaced along the θ proficiency scale from -3.0 to 3.0 in increments of 0.1. These proficiency values were considered the examinees' true proficiency. While this was not a realistic ability distribution for most testing programs it provides a reasonable number of examinees at the extremes of the distribution so that the accuracy of each test design could be evaluated along the proficiency scale. In this study, the item statistics were pre-determined and, therefore, the examinees' responses did not affect the item statistics.

3.7 Routing and Scoring

Each MST design had its own set of cut points for routing, as shown in Table 3.3 which were used to route examinees to the appropriate module in subsequent stages. The cut points were chosen so that the each final stage would roughly cover the same distance along the θ proficiency scale.

Table 3.3 Routing Cut Points on θ Proficiency Sscale for Each MST Design.

	Routing Cut Points on θ Scale	
	After Stage 1	After Stage 2
2-Stage Designs		
1-2	0.0	NA
1-3	-1.0, 1.0	NA
3-Stage Designs		
1-2-4	0.0	-1.0, 0, 1.0
1-3-5	-1.0, 1.0	-1.5, -0.5, 0.5, 1.5

The routing cut points on the θ proficiency scale were then converted to a corresponding number of correct items for an examinee with a proficiency level equal to the θ cut point. For example, the 2-stage test shown in Figure 3.4 needed a cut point to route examinees to module B or C depending on their module A performance.

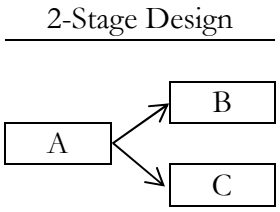


Figure 3.4 Number of Modules in the 2-Stage Design

Suppose that the routing cut point on the θ proficiency scale is 0.0. The expected number of correct answers in module A for an examinee with a proficiency of 0.0 was computed based on the item statistics for each item in module A. For an item clone, the predicted item statistics were used. As described above, the predicted statistics were the only set of item statistics known to the test administrators at the time of administration. For an examinee with a proficiency of θ, the expected number of correct responses to a set of items is the

sum of the individual probabilities of correctly responding to each item at the given proficiency level:

$$E\left(\text{number correct} \mid \theta_j\right) = \sum_{i=1}^{n} P_i\left(\theta_j\right), \tag{3.2}$$

where θ_j is the proficiency level of interest, P_i is the probability of correctly responding to item *i*, and there are *n* total items. The probability of correctly responding to each item will be based on the 3PL given in (3.1). Suppose that the expected number correct for an examinee with a proficiency of θ=0.0 is 12.3, for example. Examinees who correctly respond to 13 or more items will be routed to module C in the second stage while examinees who correctly respond to 12 or fewer items will be routed to module B.

Both routing and scoring were based on number correct rather than an examinee's estimated θ value because number correct is often used operationally for several reasons. First, it is easier to program the routing algorithms for the CAT using number correct than IRT scoring. Secondly, number correct scoring is sufficient in most cases (Luecht & Nungester, 1998) and is easier to explain to a non-technical audience. The examinee's final score on the θ proficiency scale was based on the total number of items correct across all stages. Using the procedure described earlier to determine the expected number correct for a given proficiency level, a conversion chart from number correct to proficiency on the θ scale was constructed. While examinees were simulated to have proficiencies from -3 to 3, the assigned scores ranged from -4 to 4. Any score estimates that would have been less than -4 or greater than 4 were reported as -4 and 4, respectively. Each path through the MST had its own conversion chart because the complete set of items was unique for each path.

3.8 Evaluation Criteria

One goal of this study was to determine the accuracy with which examinees' proficiency levels were estimated under the conditions described above. Each examinee's true proficiency, used to simulate responses, was compared to the examinee's estimated proficiency resulting from the simulated MST. The standard error of examinee proficiency estimates conditioned on each of the 61 true proficiency values along the θ scale.

Even if the overall standard error associated with a specific test design was relatively small, there was the possibility that a systematic error, or bias, was present. To quantify the notion of systematic error, the deviation between the estimates and truth was calculated, but with bias the direction of the error will not be ignored. This measure of bias is simply the sum of the deviations:

$$\text{Bias} = \frac{\sum_{j=1}^{J}\left(\hat{\theta}_j - \theta_j\right)}{J}, \tag{3.4}$$

where $\hat{\theta}_j$ and θ_j are the estimated and true proficiencies for examinee *j*, respectively, and *J* is the total number of examinees. The calculation of bias allows deviations in opposite directions to, in effect, cancel each other out. However, the bias calculation indicates if the deviations are greater in one direction than the other.

A subset of the fourteen conditions that were most promising, in terms of smallest bias and smallest standard error among the examinees' proficiency estimates were replicated 100 times. The resulting bias, its absolute value, and standard error for each replication were averaged over the 100 replications.

CHAPTER 4

RESULTS

4.1 Introduction

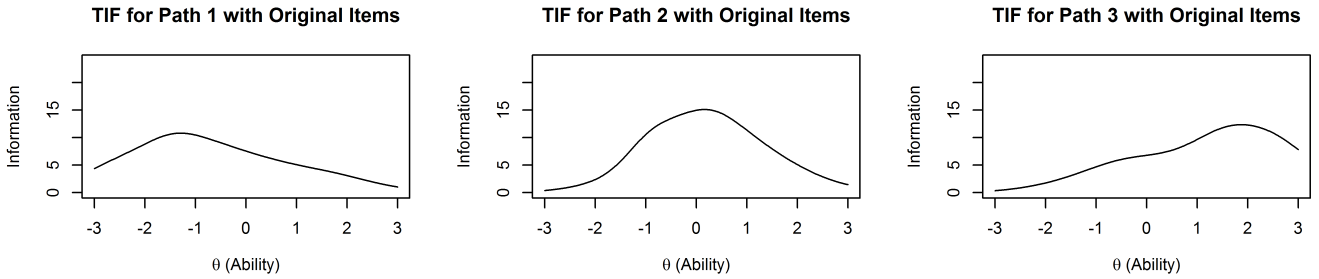
The results of the multistage test (MST) simulation are presented in this chapter. In this study, item clones were simulated using realistic assumptions based on the results of Sinharay and Johnson (2008) who administered item clones in the experimental section of the Graduate Record Examination (GRE). The goal of the item models in their study was to create a clone with the same item statistics as its parent. First, the variability introduced by the different types of item clones will be shown. Then the results of the smaller pilot study are considered and which conditions were deemed most promising based on the precision of proficiency estimates. Finally, the results of the selected conditions and their one-hundred replications will be reported.

4.2 Variability Resulting from Item Clones

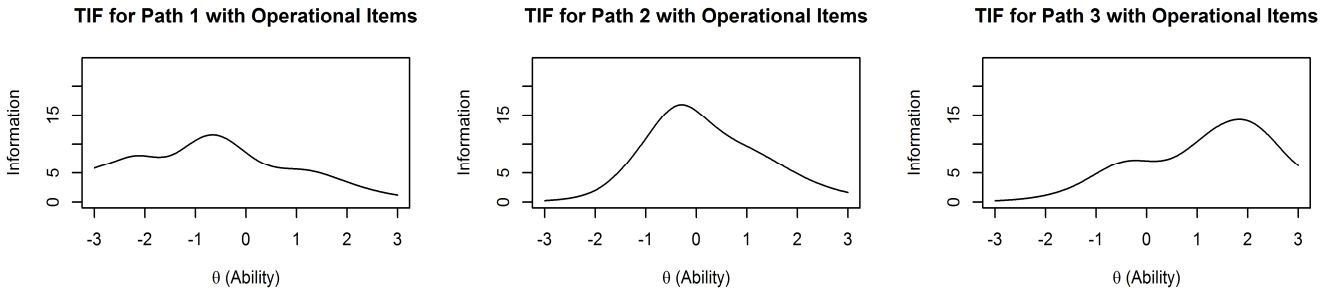
Because the results were consistent over all four test designs, two 2-stage and two 3-stage designs, only the results for the two-stage design with one routing test and three modules in the second stage, the 1-3 design, will be discussed in detail in this section. (The corresponding results for the other three designs are presented in Appendices B-D.) Across all conditions, the mean difficulties for each of the four modules in the 1-3 design were almost identical to the mean difficulties for the test with the parent items, the baseline condition. This was reasonable considering how the clones were simulated. A random number within the boundaries of the small, moderate, or large guidelines was added to the parent item's difficulty.

In Figure 4.1 the resulting test information function (TIF) for each of the three paths in the 1-3 design are presented, where path 1 is for low performers, path 2 for medium, and path 3 for high performers.

Original Items – No Cloned Items



100% of Items are Cloned with Large Variability



100% of Items are Cloned with Small Variability

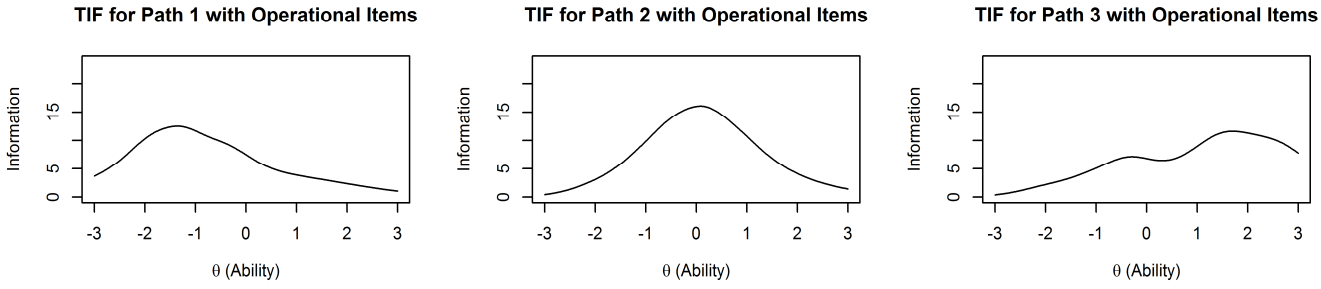


Figure 4.1 TIFs for 3 Possible Complete Paths through 1-3 Test Design for 3 Conditions

The top row of the figure shows the TIFs for the baseline condition, that is, when the original, or parent, items were used. The middle row has the TIFs when all items were clones simulated with large variability. The bottom row shows TIFs when all items are clones simulated with small variability. Even when all items were cloned and the item difficulties of the clones were simulated to vary largely when compared with the parent items the TIFs for the 3 paths do not differ greatly from the original.

While the TIFs did not indicate large variation across the three different clone conditions, the bias with which examinees' abilities were estimated did vary considerably over different conditions. The biases of the estimates resulting from three different conditions are shown in Figure 4.2. These replications used the sets of items and items clones that created the TIFs in Figure 4.1.

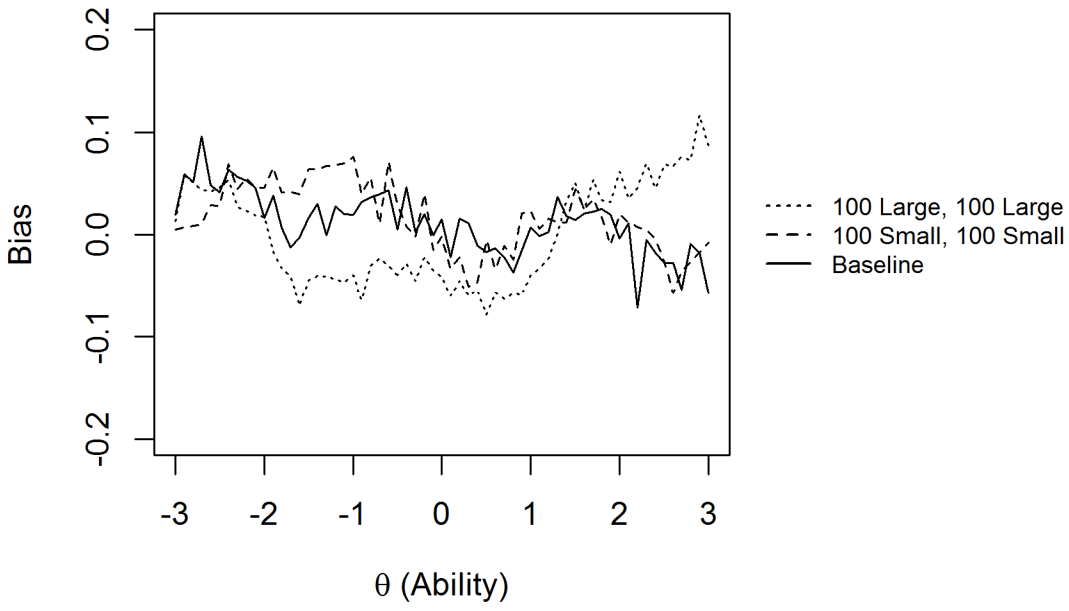


Figure 4.2 Bias of Ability Estimates for 3 Conditions in Test Design 1-3

The bias of the estimates when all items were cloned and simulated to have small variability was similar to that of the baseline condition with no cloned items. When all items were

cloned and simulated to have large variability, the ability estimates were more erratic, as can be seen in Figure 4.2. However, there were replications when all items were cloned with large variability and the resulting bias was more similar to that of the baseline condition while the clones with small variability performed more erratically, as shown in Figure 4.3. This behavior was an indication that many replications were needed of each condition in each test design.

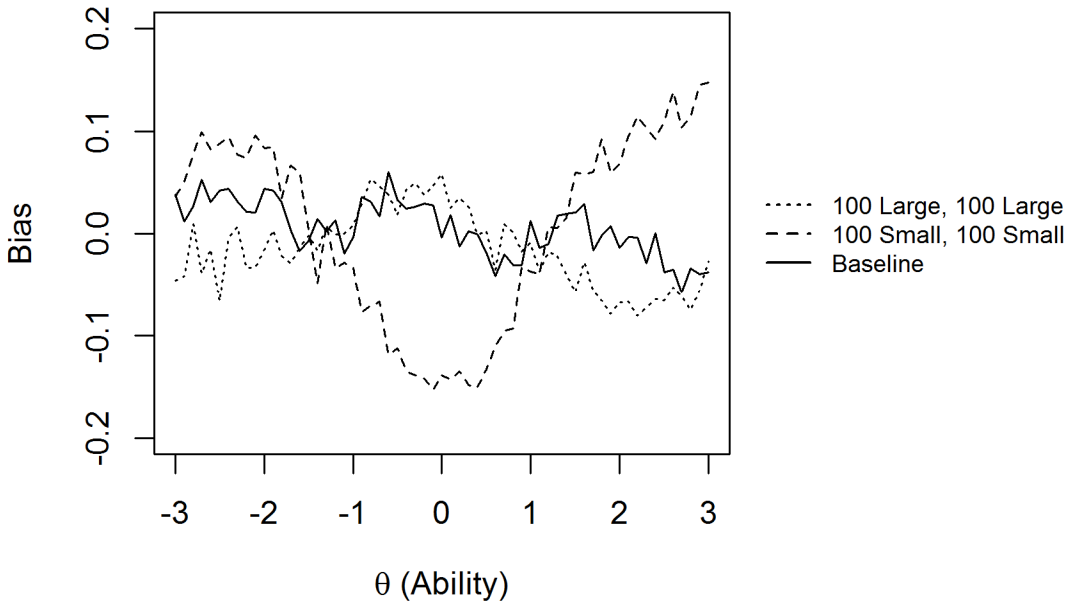


Figure 4.3 Bias of Ability Estimates for 3 Conditions in Test Design 1-3

4.3 Results of Pilot Study

Before running the 100 replications of each condition, a small pilot study was conducted with each of the four test designs and all 14 clone conditions, shown in Table 3.1. After considering the results from 5 replications, the clone conditions that were most promising, with respect to bias and accuracy, were selected for the main study. Ultimately seven clone conditions plus the baseline condition of no cloned items were chosen for

further investigation. These eight conditions in total were replicated 100 times for each of the four test designs.

The results of the 1-3 test design pilot study were similar to those of the other test designs so only the 1-3 results are presented here. The bias results for three of the more erratic conditions are shown in Figures 4.4 and 4.5.

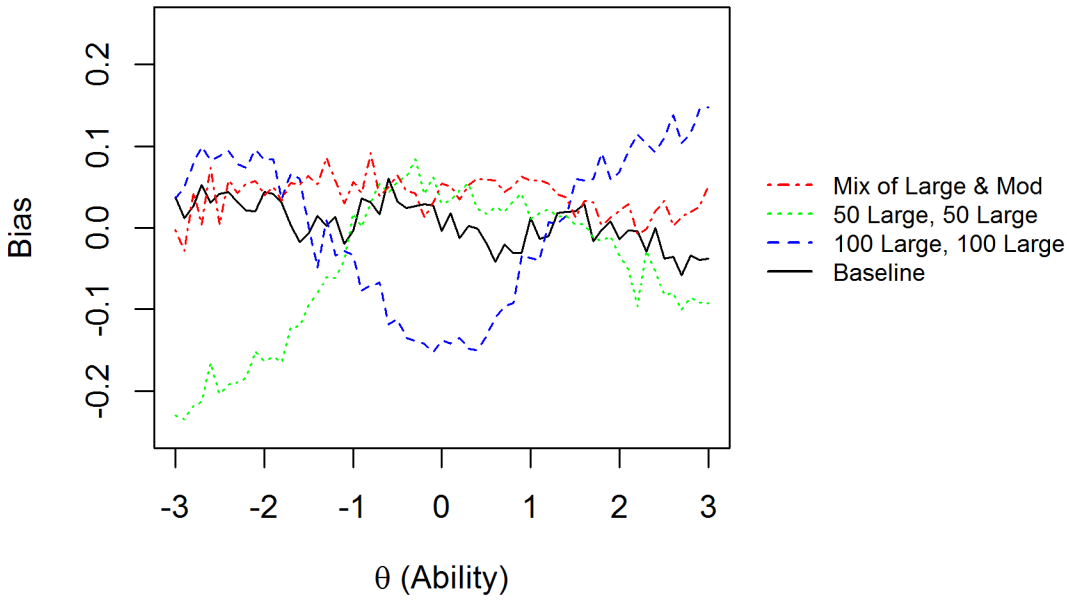


Figure 4.4 Bias of Ability Estimates for 3 Conditions in Test Design 1-3

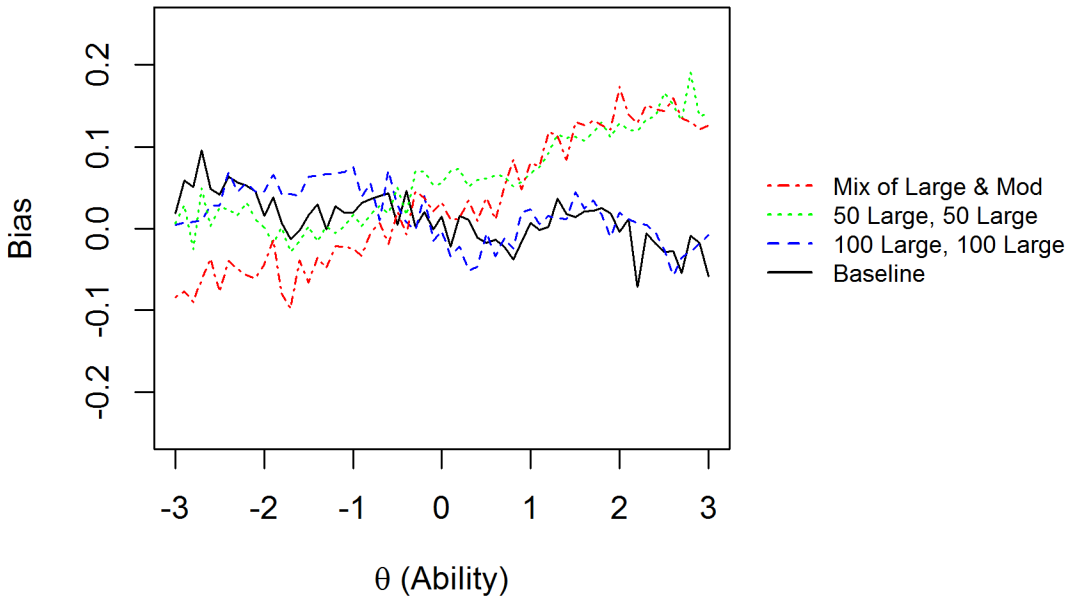


Figure 4.5 Bias of Ability Estimates for 3 Conditions in Test Design 1-3

As expected, the most erratic conditions, when compared to the baseline, involved many clones and clones with large variability. Figure 4.5 shows an example of a replication in which the condition of all items cloned with large variability had almost the same amount of bias as the baseline condition. However, for most other replications the condition with clones with large variability demonstrated considerable bias, as seen in Figure 4.4. The conditions with clones with large variability in the first stages and moderate variability for the last stage did not perform well either. The most erratic conditions were the same in all four test designs.

The results from these initial replications led to the selection of seven conditions, in addition to the baseline condition with no clones, for the main study. The seven conditions

not selected all involved moderate and large clones as half or all of the items across all stages. Of course, the most promising conditions involved clones with small variability; for that reason three conditions were selected: one-third, one-half, and all cloned items with small variability. The next two conditions included moderately and largely variable clones: one-half of items cloned with moderate variability and one-third of items cloned with large variability. Finally, the combination of moderately variable clones in the first stage and no clones in subsequent stages was considered. Specifically, two conditions were evaluated for this combination: all first stage items were clones and exactly half of first stage items were clones. The complete set of conditions is presented in Table 4.1 along with their associated abbreviations that will be used in graphs and tables.

Table 4.1 Main Simulation Study Conditions for Two-Stage Test Designs

Description of Condition	Abbreviation
Baseline, no item clones	No Clones, No Clones
One-third of items cloned with small variability	33 Small, 33 Small
One-half of items cloned with small variability	50 Small, 50 Small
All items cloned with small variability	100 Small, 100 Small
One-half of items cloned with moderate variability	50 Mod., 50 Mod.
One-third of items cloned with large variability	33 Large, 33 Large
One-half of first stage items cloned with moderate variability	50 Mod., No Clones
All first stage items cloned with moderate variability	100 Mod., No Clones

4.4 Results of Main Simulation Study

The results for both the two-stage and three-stage test designs were consistent with respect to the conditional standard errors and the bias of the proficiency estimates, therefore only the results of the two-stage 1-3 test design will be presented here. (The corresponding graphs for the other three test designs are provided in Appendices B-D.) First, consider the

resulting conditional bias associated with each condition. The conditional bias was calculated for each for each of the 100 replications then the mean bias for the 100 replications was calculated at 61 ability levels evenly spaced from -3.0 to 3.0 on the theta ability scale. The absolute bias was used to remove the effect of "canceling" across the one-hundred replications. Figure 4.6 shows the mean absolute bias for test design 1-3 when one-third of the items are cloned with small variability and when one-third of the items are cloned with large variability.

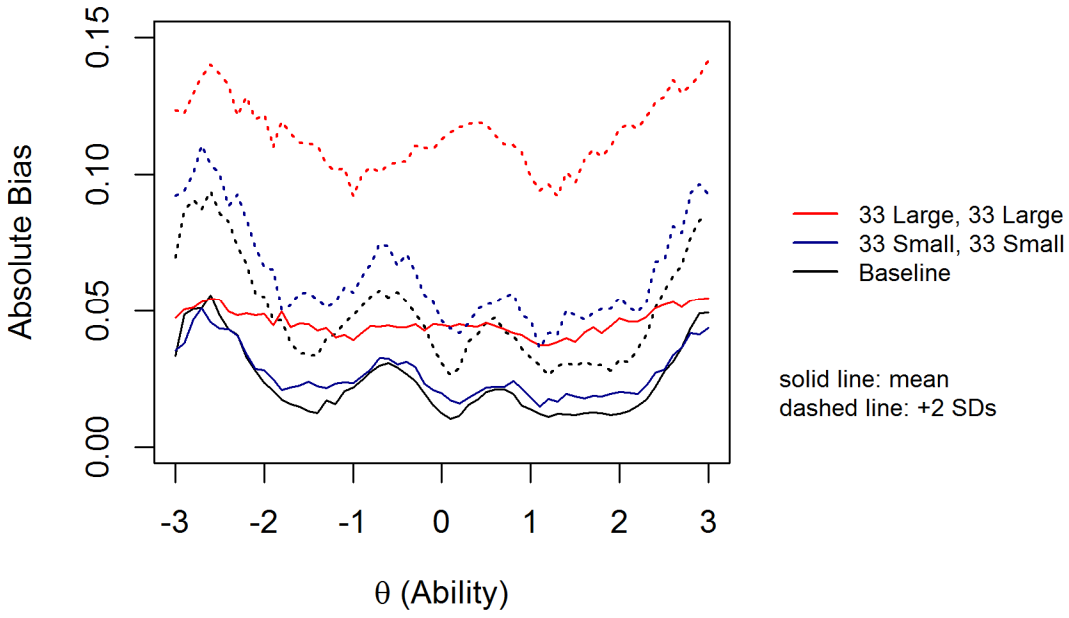


Figure 4.6 Mean Absolute Bias over 100 Replications for 1-3 Test Design

The group of lines at the bottom of the graph represents the mean absolute bias for the two conditions plus the baseline and the pair of lines at the top of the graph are error bands indicating the one-sided confidence interval for the mean absolute bias. Figure 4.7 displays the mean standard errors over the one-hundred replications for same conditions, one-third clones with small and with large variability.

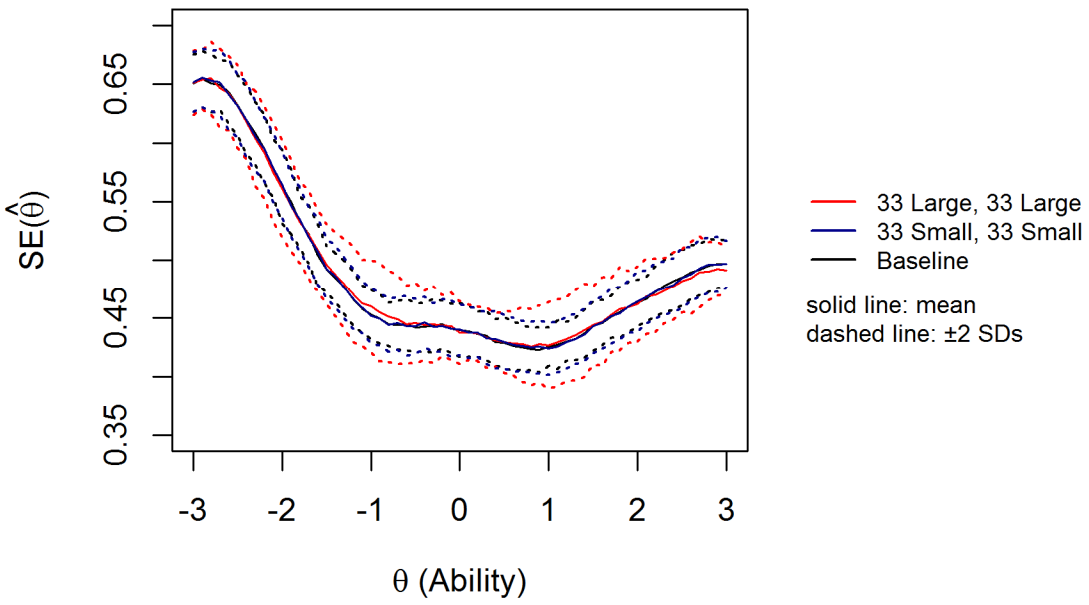


Figure 4.7 Mean Standard Error over 100 Replications for Test Design 1-3

It is apparent that the error bands are wider when the clones have more variability and the mean bias more closely follows the baseline condition when there is only small variation in the clones.

Rather than inspect the conditions in pairs, Figure 4.8 displays the mean absolute bias and error bands for three of the seven conditions: one-third, one-half, and all of the items cloned with small variability in both stages.

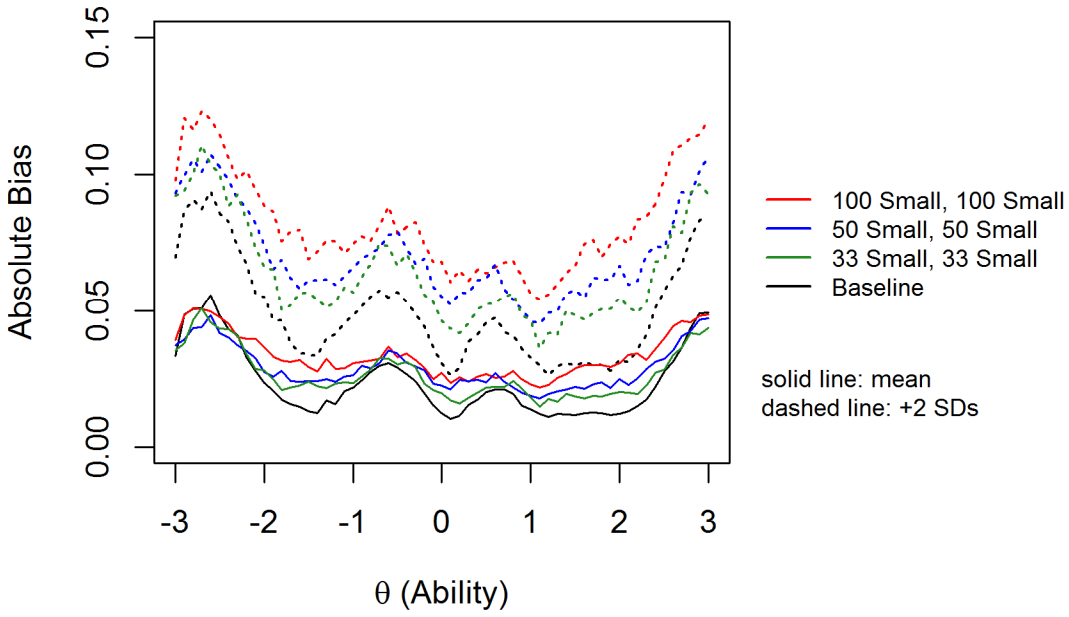


Figure 4.8 Mean Absolute Bias over 100 Rreplications and One-Sided Error Bands for 1-3 Design when Item Clones have Small Variability

These three conditions had the narrowest error bands of all seven conditions and closely followed the mean absolute bias of the baseline condition with no clones. The conditional standard errors for these three conditions are shown in Figure 4.9.

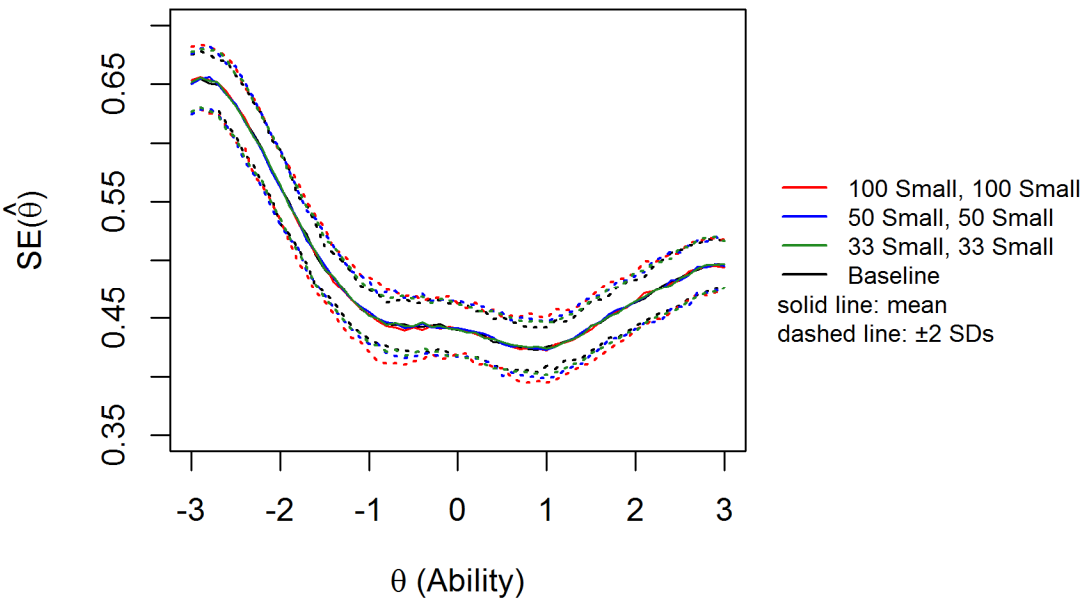


Figure 4.9 Mean Standard Errors over 100 Replications and Confidence Bands for 1-3 Design

The range of the standard error was noticeably wider around θ equal to -1 and 1 when all of the items were cloned than when only one-third or half of the items were cloned.

As expected, the conditions with clones with small variability performed most closely to the baseline condition of no clones. Figures 4.10 and 4.11 display the mean absolute bias and standard error for the other four conditions: one-third item clones with large variability, one-half item clones with moderate variability, and one-half and all first stage items with moderate variability.

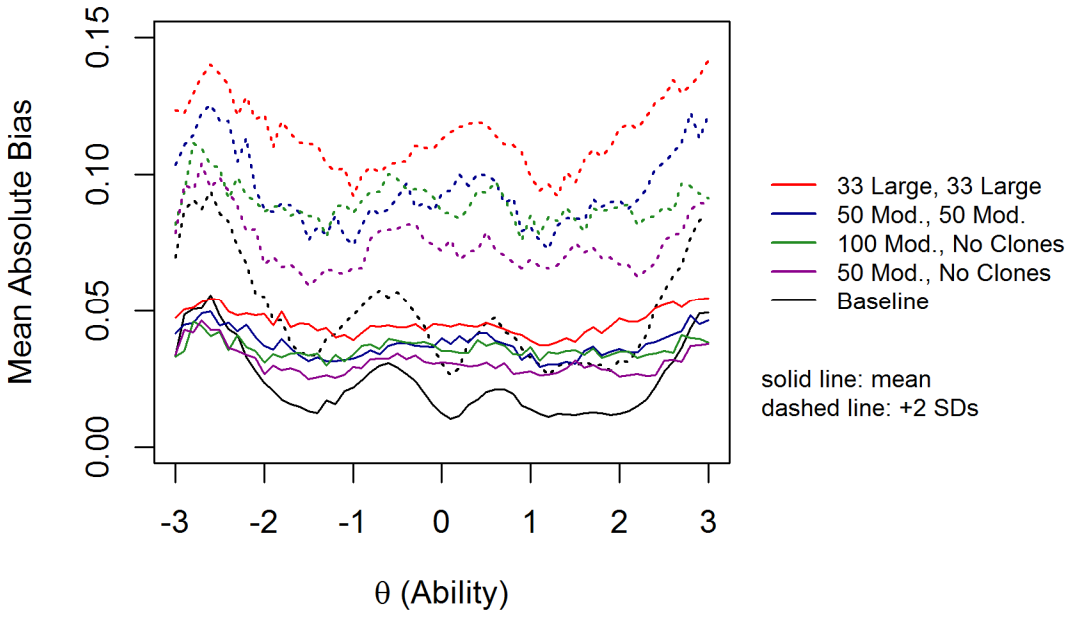


Figure 4.10 Mean Absolute Bias for Test Design 1-3

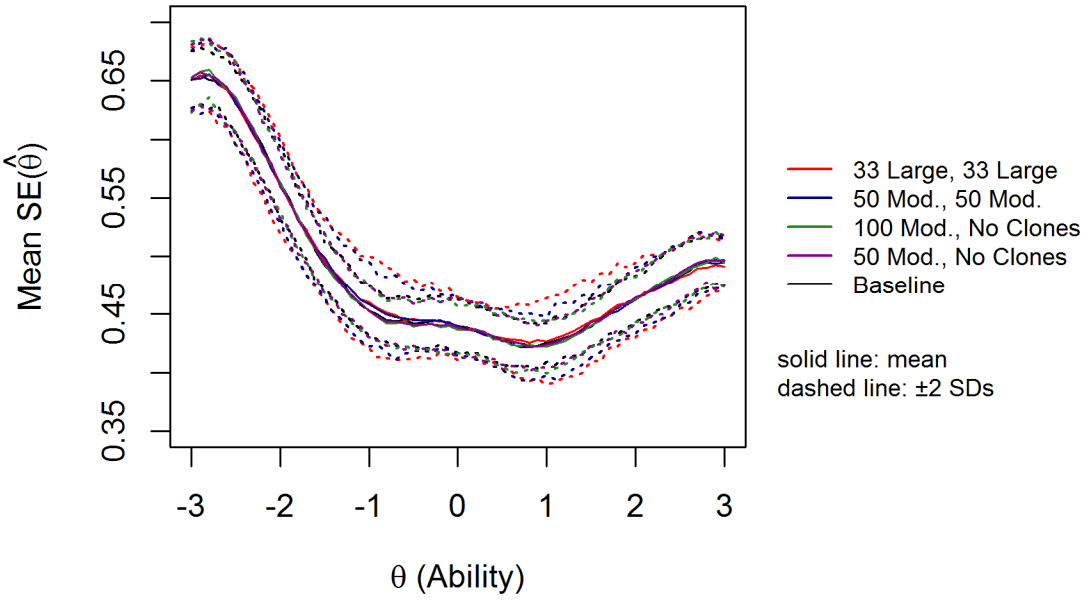


Figure 4.11 Mean Standard Errors for Test Design 1-3

Again, the results were consistent with expectations that the more items cloned and the more the clones vary from their parent items, the more bias the ability estimates had. Specifically, the most promising condition in this second group was when only half of the items in the first stage were cloned and none of the items in the second stage. This condition did not perform as well as the conditions with small variability in one-third or even half of the items, when each is compared to the baseline conditions, see Figure 4.12.

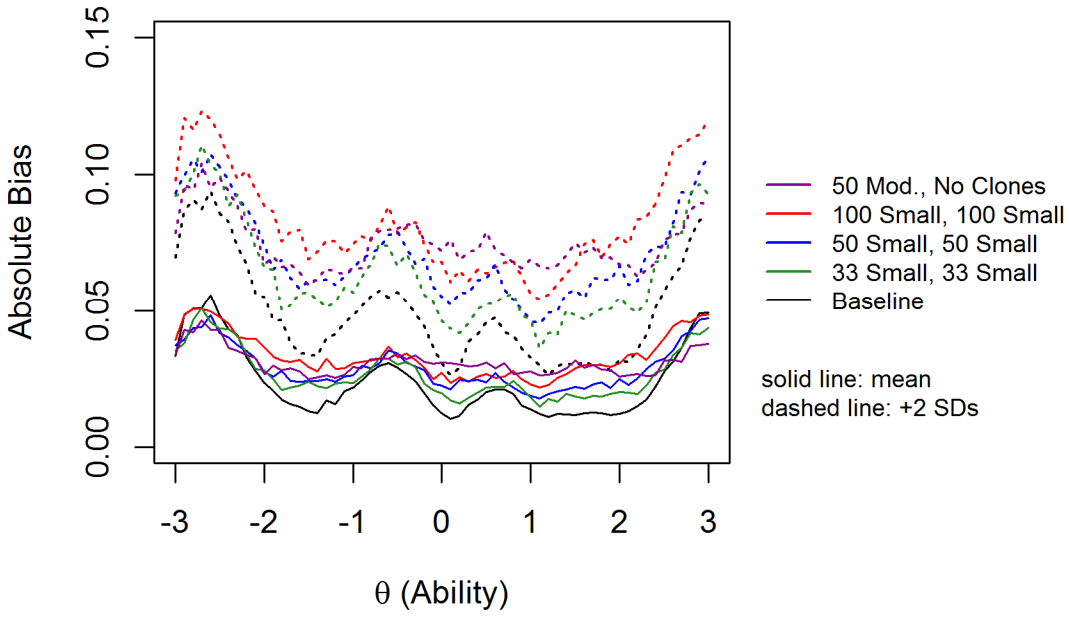


Figure 4.12 Mean Absolute Bias for Test Design 1-3 for 4 Conditions

4.5 Summary of Results

The results of each condition with respect to the accuracy and variability of the ability estimates were consistent over the four test designs. Across all four test designs the more variable the simulated clones and the more clones per test, the more biased and variable the examinees' proficiency estimates were, as was to be expected. If the variability and bias associated with cloning one-third of the items when the clones varied very little from their parents' item statistics are too much for a particular testing program, then it is clear that other conditions will also not suit that program's needs. Rather than focusing on which condition was better than the other, because, those results were not surprising, the next chapter will consider which level of cloning would be adequate for different needs. To address this, the variability and bias associated with the different conditions will be quantified with respect to a hypothetical scaled score in the next section.

The results of this study have shown that unless the item model can accurately predict the behavior of its clones, the ability estimates can be quite variable. In the work presented here, "small variability" in the item clones was simulated based on the results of one previous study, however, those interested in a particular item model may have support for less variability between the clone and its parent, in which case the ability estimates would be less biased.

CHAPTER 5

DISCUSSION

5.1 Study Design

The results of this simulation reveal the variability/bias in the estimates of examinees' abilities when even small variability is introduced between an item clone and its parent. The same set of parent items were used for all one-hundred replications and across all conditions within each of the four test designs. Additionally, 1000 examinees were simulated at each of the 61 ability levels along the θ scale.

The item statistics for the parent items were taken from an operational testing program. The distribution of difficulty parameters in the operational items were not ideal for creating a set of items such that the resulting test information function (TIF) would be completely flat across the ability scale, as would be optimal when attempting to provide all examinees with estimates of their ability with equivalent errors of measurement regardless of the ability. Had more extreme items, both easy and difficult, been available, it would have been more likely to select items in the later stage modules such that the TIF would be more consistent for all examinees. This inability to select optimal sets of items per module per stage leads to the fluctuation in precision along the ability scale. The reduced precision at the extremes of the ability scale can be seen in Figure 5.1.

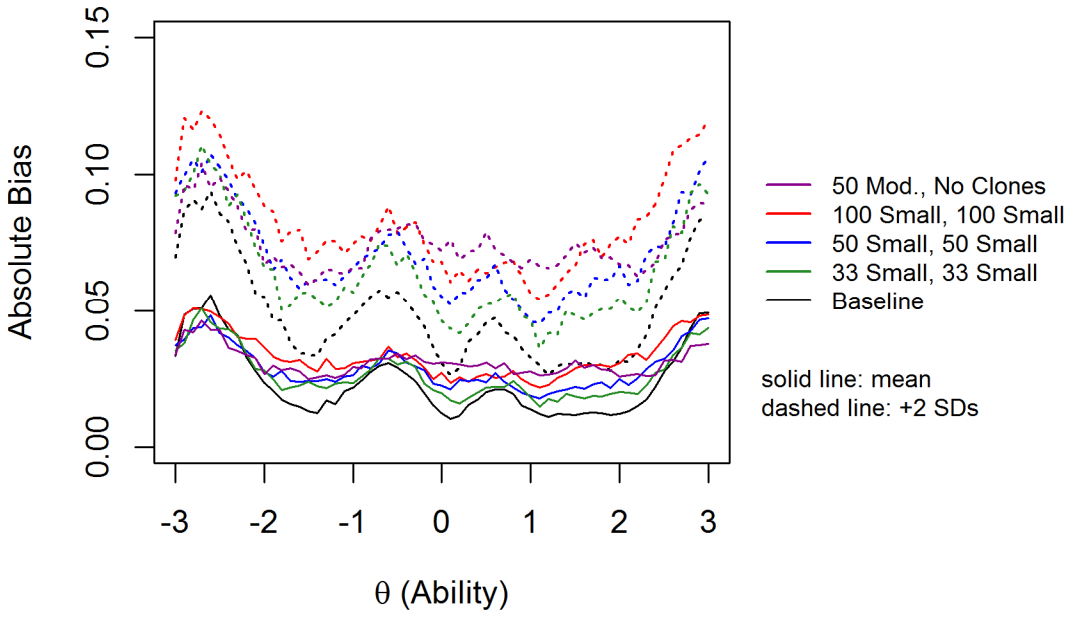


Figure 5.1 Mean Absolute Bias for Test Design 1-3 for 4 Conditions

Even in the baseline conditions with no cloned items, the bias changes along the θ scale. This fluctuation was consistent across the different clone conditions and does not affect the interpretation of these simulated results. However, an operational testing program interested in accurately measuring the examinees all across the ability scale, not just at particular cut scores, for example, would want to address this fluctuation by selecting more appropriate sets of items, if possible.

5.2 Four Test Designs

Two 2-stage and two 3-stage test designs were studied. As reported in the previous section, the resulting bias and accuracy of examinees' ability estimates based on the various conditions were consistent across the four test designs. Specifically, the ordering of the conditions with respect to less bias, more accuracy, was consistent across designs. However,

the degree to which one condition was more or less accurate or biased than another condition did vary across test designs. This difference was most apparent in the conditions with clones in only the first stage. This difference is reasonable when the test designs are considered. All tests had a total of 36 items, while the 2-stage design had 18 items per level and the 3-stage design had 12 items per level. For example, the scenario in which half of the items in the first stage are cloned to have moderate variability shows less bias in the 3-stage design than in the 2-stage design.

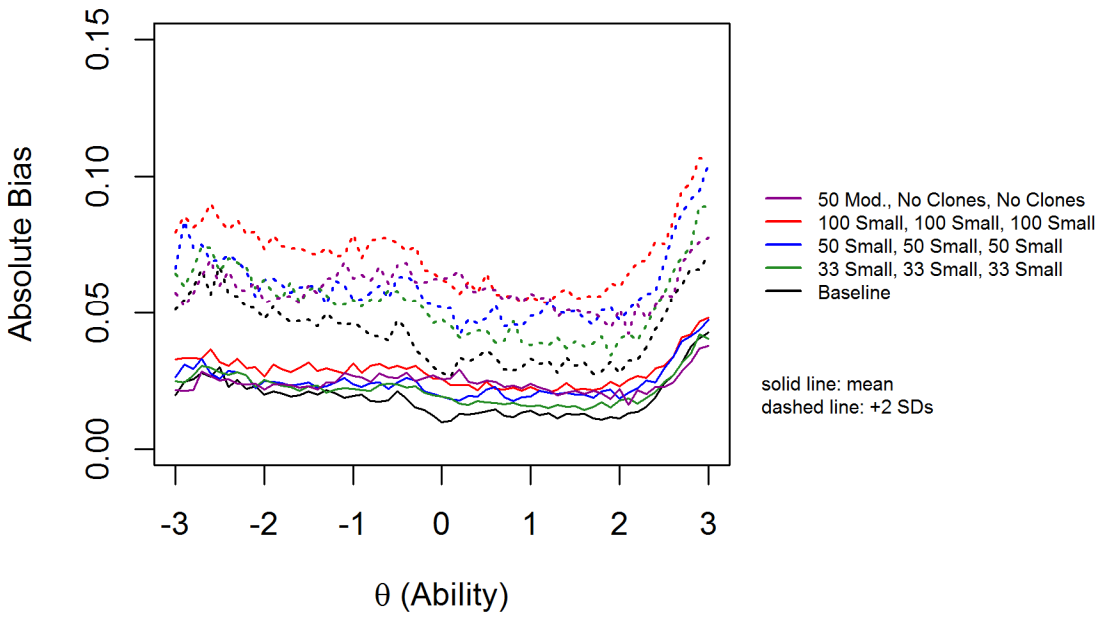


Figure 5.2 Mean Absolute Bias for Test Design 1-3-5 for 4 Conditions

In the 3-stage 1-3-5 design, the conditions with 50% moderate clones in the first stage only, indicated by purple in Figure 5.2 ("50 Mod., No Clones, No Clones"), performed better, almost everywhere along the θ scale, than the condition when all items were cloned and simulated to have small variability, indicated by red. However, this condition with 50% moderate clones in the first stage only performed similarly to the 100% small clones

condition in the 2-stage 1-3 design as seen in Figure 5.1. Of course, the obvious reason for this difference is that only 12 items were cloned in the 3-stage design, whereas 18 items were cloned in the 2-stage design, thus introducing more variability.

5.3 Hypothetical Scaled-Score

To better illustrate the effects of the cloned items in a multistage test (MST), a hypothetical scaled-score will be used for the 2-stage design with one routing test in the first stage and three modules in the second stage, the 1-3 design. Suppose that from -3.0 to 3.0 on the θ scale is mapped to a scaled-score of 200 to 800. For examinees whose true score falls between 300 and 700, the absolute bias they would have experienced would have been less than about 6 points on the 200-800 scale, as shown in Figure 5.3. If all of the items were cloned and simulated to have small variability from the parent items, the majority of examinees would have less than about 10 points of absolute bias on the 200-800 scale (and less than 0.7 points on a 130-170 point scale).

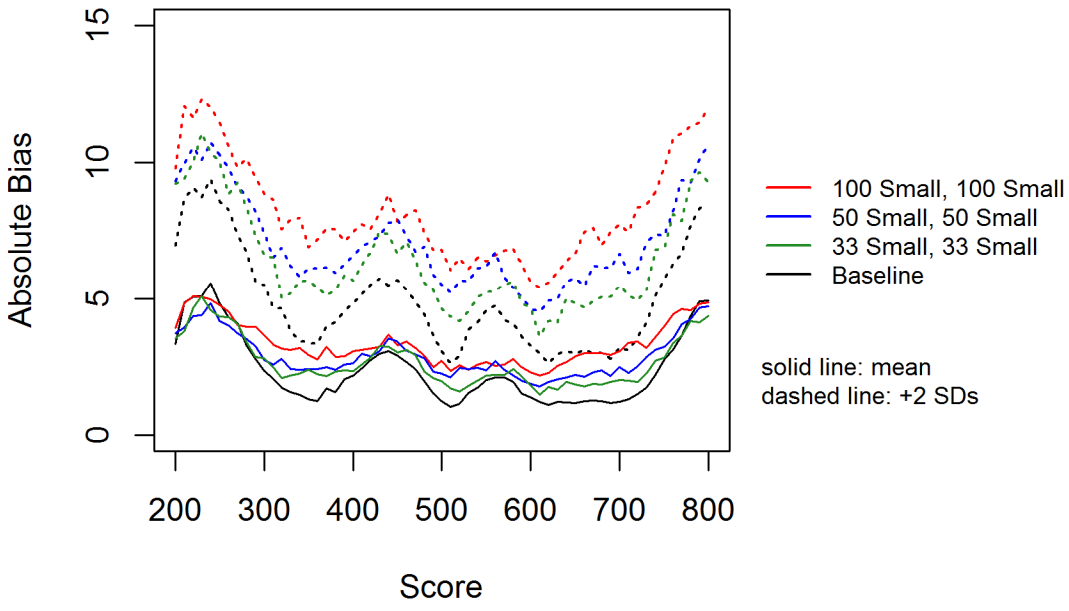


Figure 5.3 Mean Absolute Bias for Test Design 1-3 on 200-800 Point Scale

Even if half of the items in the first stage were simulated to be clones with moderate variability when compared to the parent items, the mean absolute bias would have been less than about 10 score points, for almost all examinees, on the 200-800 point scale. If this level of absolute bias is acceptable to a particular testing program, then automatic item generation with the goal of creating clones which are only moderately similar to the parent items is feasible.

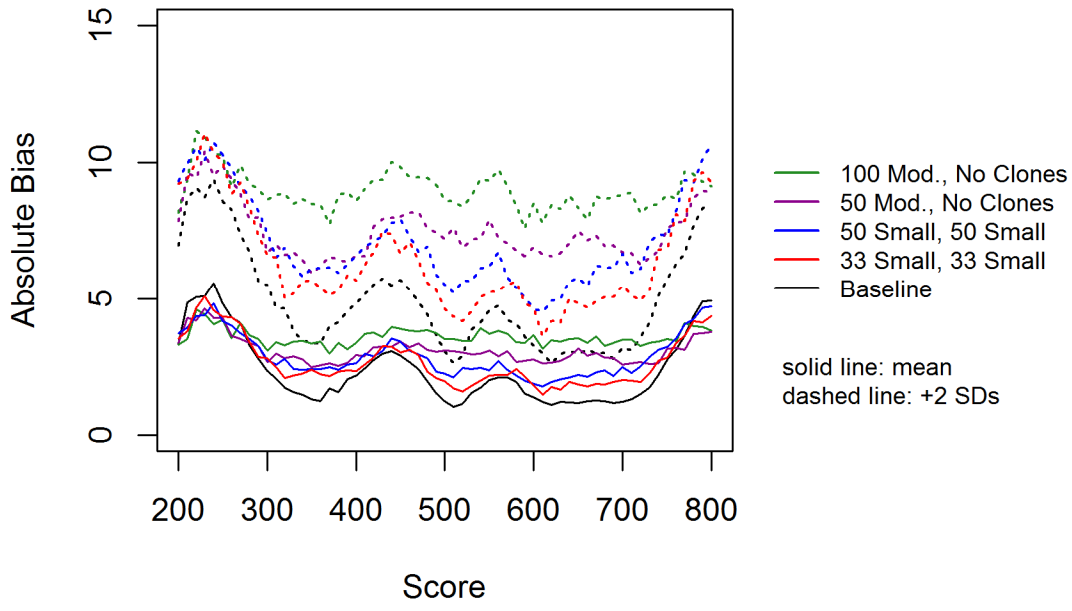


Figure 5.4 Mean Absolute Bias for Test Design 1-3 on 200-800 Point Scale

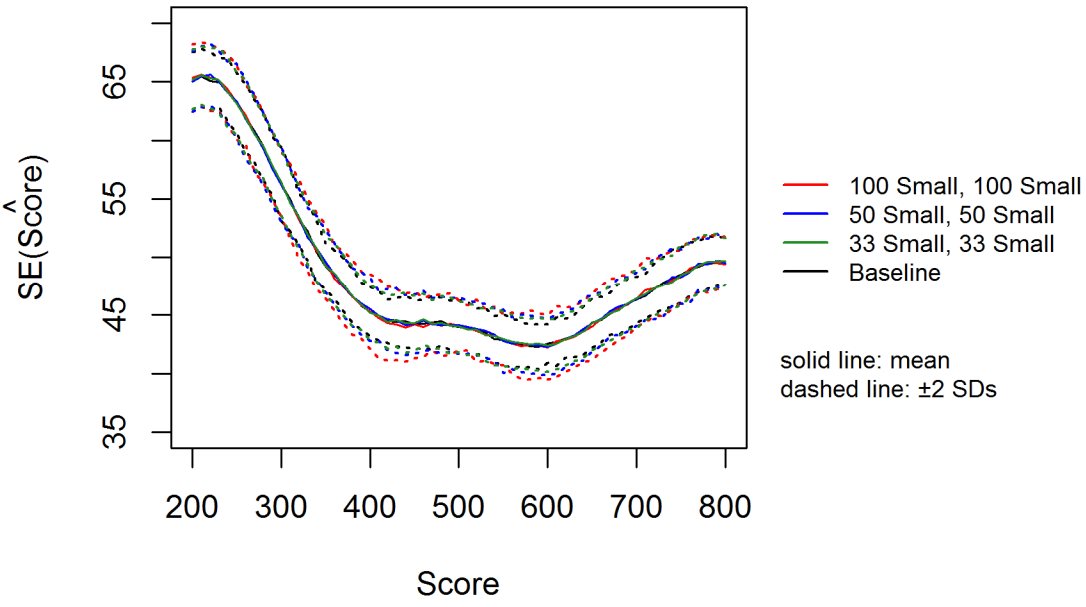


Figure 5.5 Mean Standard Error for Test Design 1-3 on 200-800 Point Scale

5.4 Variability Across Replications

Because the same set of parent items was used for a given test design, the only variability was in the random performance of the examinees and the clones that were randomly generated according to the rules of the various conditions studied. It was striking how much variability occurred within one condition with the same set of parent items. The bias for a few of the replications was similar to the bias when no clones were used, but there were replications that deviated considerably. In Figure 5.6 the bias for Test Design 1-3 is shown for three replications when all items were cloned with small variability and the condition of no clones. The bias for examinees of average ability was in the opposite direction when compared with the two other replications and when no clones were used, while for the upper ability levels two of the replications had considerably larger bias.

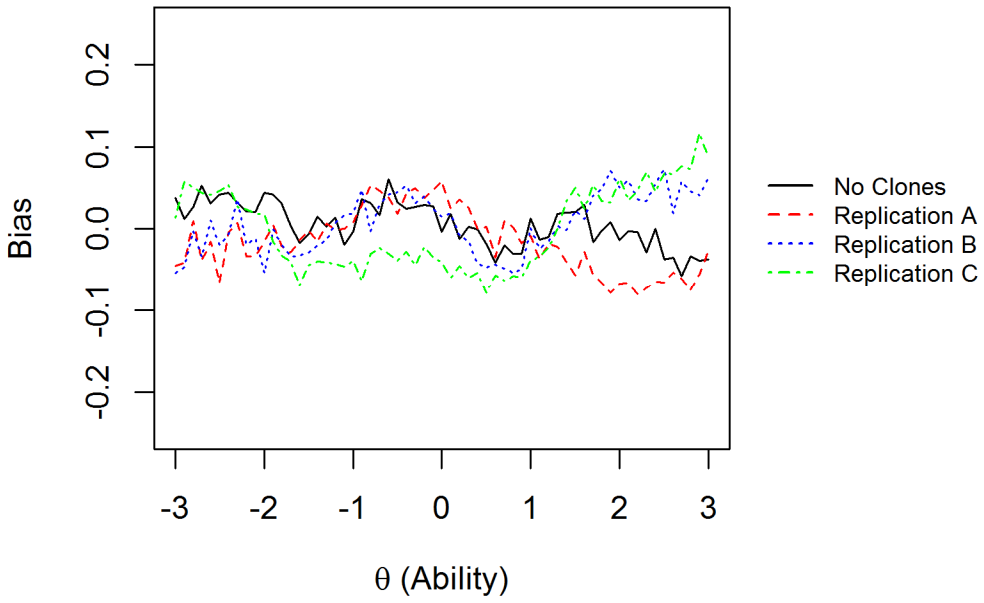


Figure 5.6 Three Replications with 100% Clones of Small Variability for Test Design 1-3

Even if the simulated variability was small and not all of the items were cloned, the resulting bias and accuracy varied considerably as can be seen in Figure 5.7 when half of the items were cloned and simulated to have small variability. One replication, labeled replication B on the graph, had more bias than the other replications and the no clones condition for the lower half of the ability scale, while the difference among the replications was smaller for the higher ability levels.

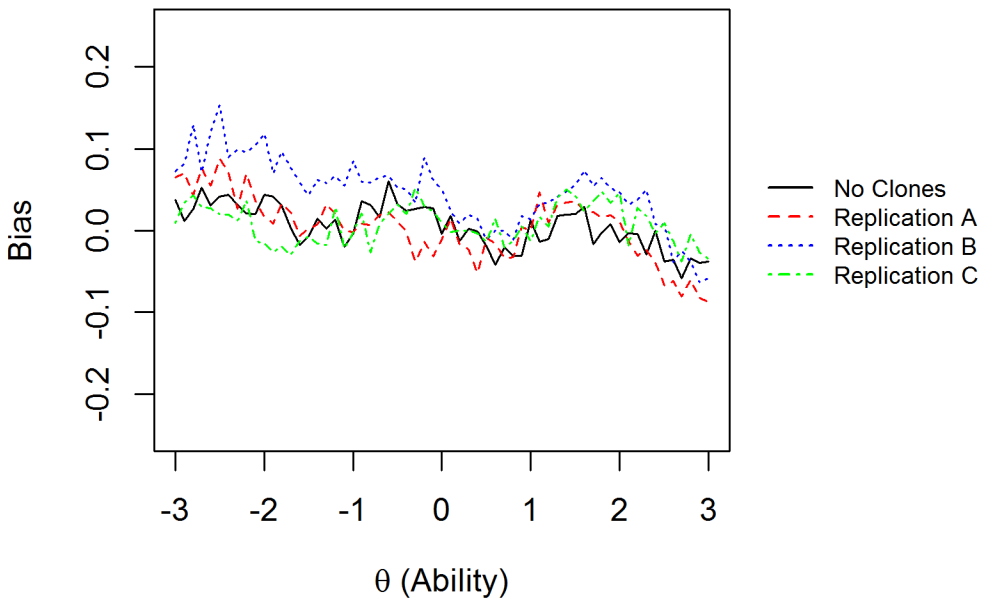


Figure 5.7 Three Replications with 50% Clones of Small Variability for Test Design 1-3

Even though the mean absolute bias over one-hundred replications may appear well-behaved and similar to the baseline condition when no clones are used, it is important to note that the possibility of a particular set of clones behaving erratically is possible and could lead to less than desirable bias and accuracy of examinees' proficiency estimates.

5.5 Future Research

This study was based on the assumption that the goal of the examination was accurately measuring examinees along the ability scale. However, it would be interesting to consider the implication of automatic item generation in an MST in which the test developers are only interested in categorizing examinees, such as basic, proficient, and

advanced, or even, simply, as passing or failing, as in a credentialing exam. If the desire for greater measurement precision is located at only a few points along the ability scale, then one could consider using clones only if the item difficulty is a certain distance away from the location of the cut points, for example.

This study used number-right scoring, rather than item response theory (IRT) based scoring, both for routing examinees and for providing the final ability estimates. Because number-right scoring is less computationally intensive it is reasonable to use it for routing examinees through the test, but the final scoring of the examinees could be done using IRT-based scoring. While this could offer a slightly more accurate scoring of examinees, the magnitude of the improvement would need to be studied. Additionally, an investigation into the relationship of test length and use of clones would demonstrate how many additional items would be needed to maintain a specified level of accuracy. A testing program may accept a moderate increase in test length if it could avoid pilot testing items.

If scores do not need to be reported immediately, it would not be necessary to use the hypothesized item statistics for the cloned items when scoring the examinees. The hypothesized item statistics, or item statistics of the parents items in this study, would still be used for routing, but if time allows, the automatically generated items which had not been calibrated or pre-tested could then be calibrated after test administration as in a linear test administration which conducts calibration and equating after administration.

The most intriguing future research in this area is, as one might expect, the more difficult: Developing accurate and useful item models for improved automatic item generation. The more that a testing program can generate items that behave as expected and appear different to the examinees, the more accurate the examinees' ability estimates will be with the added bonus of a reduction in item exposure.

5.6 Conclusions

The efficacy with which items cloned on-the-fly, with no time for calibration or pre-testing, can be used in a multistage test hinges upon the ability with which item statistics can be accurately modeled cloned items. The simulation in this study mimicked the degree of variation between a cloned item and its parent item found by Sinharay and Johnson (2008) in quantitative Graduate Record Examination (GRE) items in an operational setting. The current study used the item statistics of the clones to simulate examinees' responses, but used the item statistics of the parent items for routing decisions and scoring. This afforded the ability to determine how accurately an examinee's ability level can be recovered when the items administered to the examinee have unknown item statistics and the testing program can only use the item statistics of each item clone's parent item. In this simulation study the degree to which each item clone differed from its parent item was varied and was considered to have a small, moderate, or large difference. The percentage of exam items that were simulated to be item clones was also varied.

The results of the simulation study were not surprising with respect to the ordering of the results, specifically, the examinees' abilities were not as accurately estimated when the simulated conditions involving more item clones and/or item clones that varied more from the parent items. However, for a specific testing program, the difference between an examinee's true ability and estimated ability could be within an acceptable level. On a 200-800 point scale for a 2-stage test with one routing test and three modules in the second stage, the absolute bias was less than about 10 points for almost all examinees both when all items were simulated to be clones with small variation from their parent items and when all items in the first stage were simulated to be clones with moderate variation and no item clones in the second stage. Depending on the needs of the testing program, the nature of the test and

the decisions made based on the test results, this bias may be considered tolerable and thus allow the testing program to include clones or automatically generated items. If immediate reporting of results is not necessary, then the item clones can be calibrated after test administration as estimates of examinees' abilities would be even more accurate; only the routing of the examinees through the MST would depend on the hypothesized item statistics for any items that were clones.

This simulation study, in a sense, presented the worst case scenario of employing automatic item generation in an MST: All cloned items were assumed to have the same item statistics as the parent items, examinee scoring was based on the parent items' statistics rather than delaying the scoring until the new, cloned items have been calibrated, and item clones were distributed evenly along the item difficulty scale, rather than inserting clones strategically to replace items along the theta ability scale where ability estimates are more accurate. Yet, even with this, the accuracy observed in this simulation could still allow some testing programs to incorporate automatic item generation and maintain the integrity of their test. Furthermore, in this study the calibrated parameters of the non-cloned items were assumed to be without error, which is not realistic, so the impact of cloning would be even smaller when compared to the baseline condition. The potential for automatic item generation use is great, from achievement and credentialing exams to formative assessments in K-12 classrooms. The notion of on-the-fly item generation coupled with multistage testing could enable more individualized testing and instruction, hopefully engaging more students and leading to improved educational opportunities.

Appendix A

DESCRIPTIVE STATISTICS FOR ITEM PARAMETERS AND TIFS FOR ALL PATHS IN TEST DESIGNS 1-3-5, 1-2, AND 1-2-4

Table A.1 Descriptive Statistics for Item Parameters in 1-3-5 MST Design

Mean (Standard Deviation) of Item Parameters Per Module

	Mean (Standard Deviation) of Item Parameters		
Module	*a*	*b*	*c*
Stage 1	1.177 (0.22)	0.121 (1.10)	0.184 (0.06)
Stage 2 Low	0.985 (0.30)	-1.979 (0.60)	0.153 (0.09)
Stage 2 Medium	1.107 (0.22)	0.007 (0.56)	0.12 (0.10)
Stage 2 High	1.103 (0.32)	1.993 (0.49)	0.174 (0.13)
Stage 3 Low	0.909 (0.28)	-2.516 (0.30)	0.195 (0.11)
Stage 3 Med - Low	1.023 (0.22)	-1.546 (0.31)	0.177 (0.07)
Stage 3 Medium	1.117 (0.27)	0.011 (0.54)	0.173 (0.07)
Stage 3 Med - High	1.153 (0.24)	1.457 (0.31)	0.130 (0.13)
Stage 3 High	1.115 (0.26)	2.499 (0.26)	0.167 (0.13)

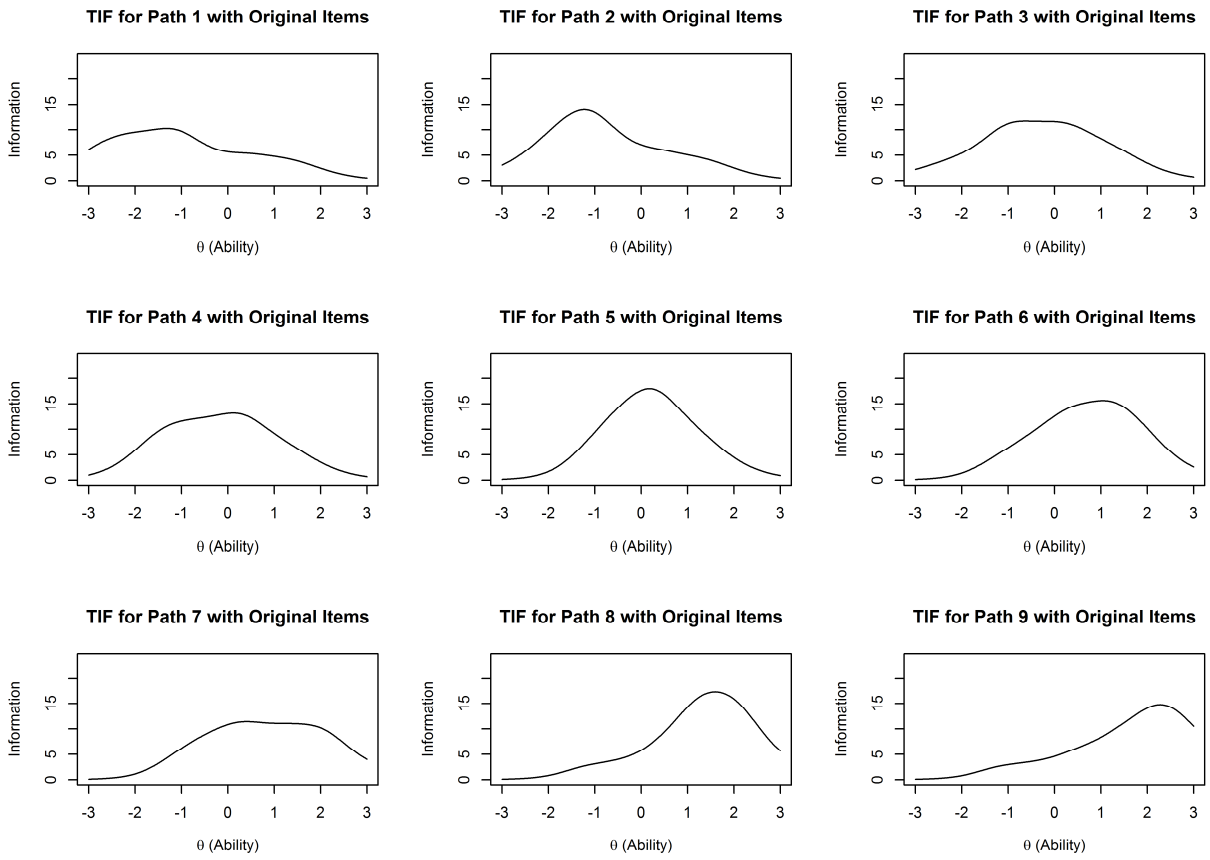


Figure A.1 TIFs for All Paths in Test Design 1-3-5

Table A.2 Descriptive Statistics for Item Parameters in 1-2 MST Design

Mean (Standard Deviation) of Item Parameters Per Module

	Mean (Standard Deviation) of Item Parameters		
Module	*a*	*b*	*c*
Stage 1	1.115 (0.21)	0.092 (1.34)	0.163 (0.06)
Stage 2 Low	1.032 (0.25)	-1.539 (0.84)	0.180 (0.06)
Stage 2 Medium	1.122 (0.30)	1.508 (0.89)	0.169 (0.12)

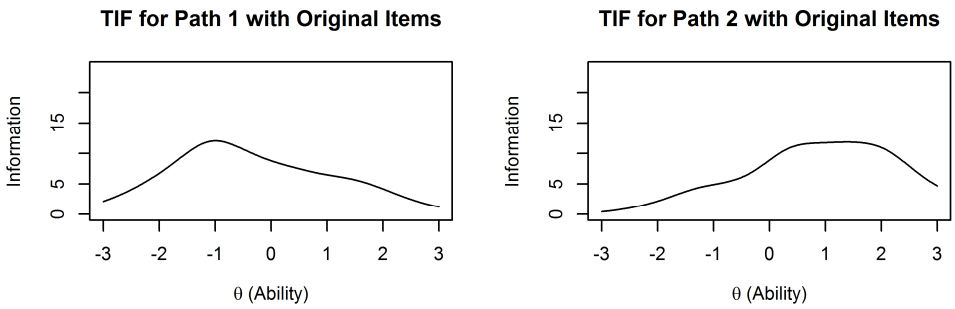


Figure A.2 TIFs for Both Paths in Test Design 1-2

Table A.3 Descriptive Statistics for Item Parameters in 1-2-4 MST Design

Mean (Standard Deviation) of Item Parameters Per Module

	Mean (Standard Deviation) of Item Parameters		
Module	*a*	*b*	*c*
Stage 1	1.177 (0.22)	0.121 (1.10)	0.184 (0.06)
Stage 2 Low	1.013 (0.28)	-1.578 (0.83)	0.183 (0.06)
Stage 2 High	1.158 (0.36)	1.476 (0.91)	0.17 (0.13)
Stage 3 Low	0.909 (0.28)	-2.516 (0.30)	0.20 (0.11)
Stage 3 Med – Low	1.128 (0.19)	-1.073 (0.58)	0.15 (0.06)
Stage 3 Med – High	1.007 (0.18)	1.042 (0.66)	0.14 (0.10)
Stage 3 High	1.115 (0.26)	2.499 (0.26)	0.17 (0.13)

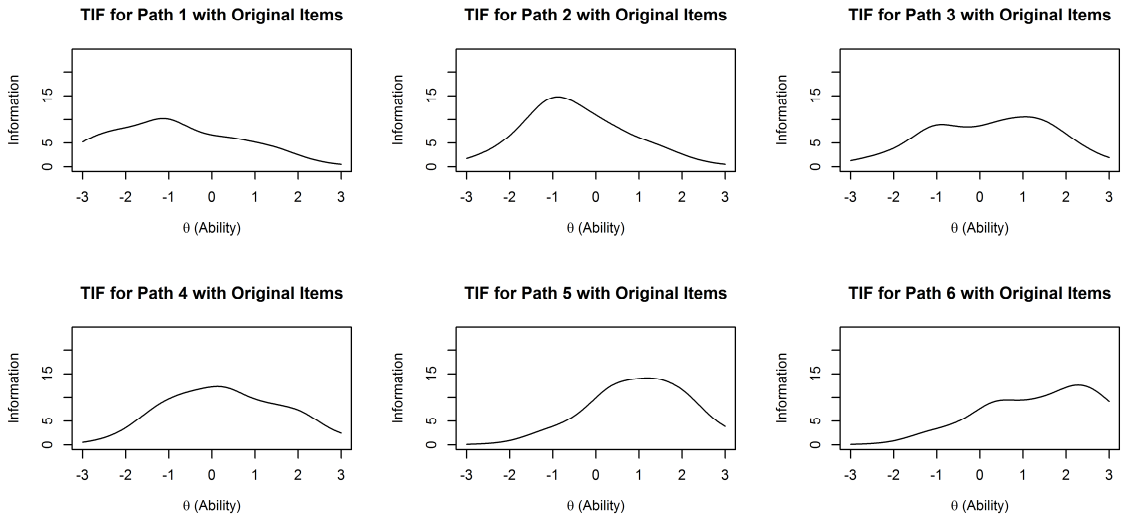


Figure A.3 TIFs for All Paths in Test Design 1-2-4

APPENDIX B

RESULTS FOR THREE-STAGE TEST: DESIGN 1-3-5

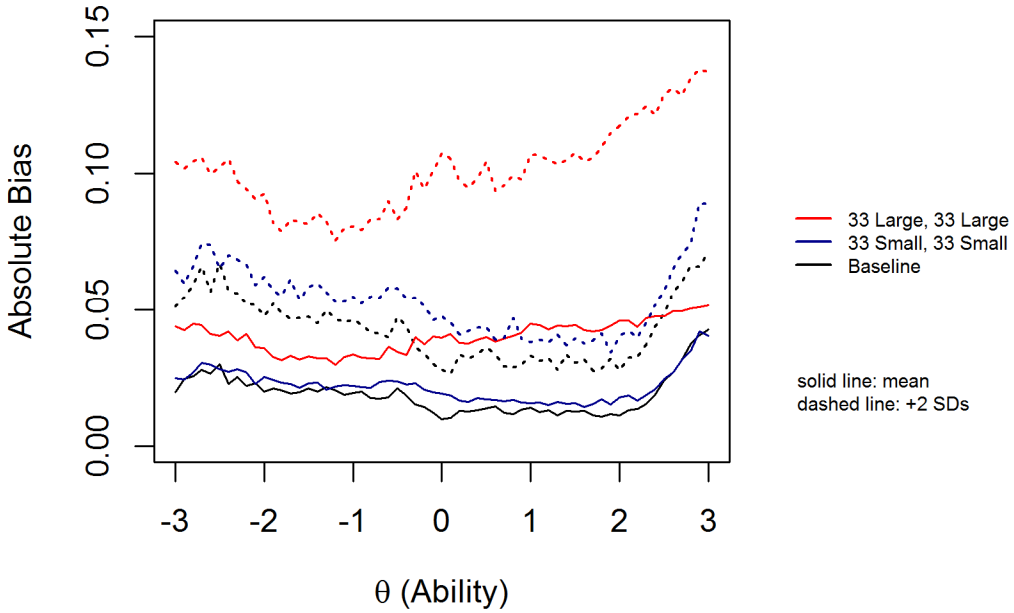


Figure B.1 Mean Absolute Bias for 2 Conditions for Test Design 1-3-5

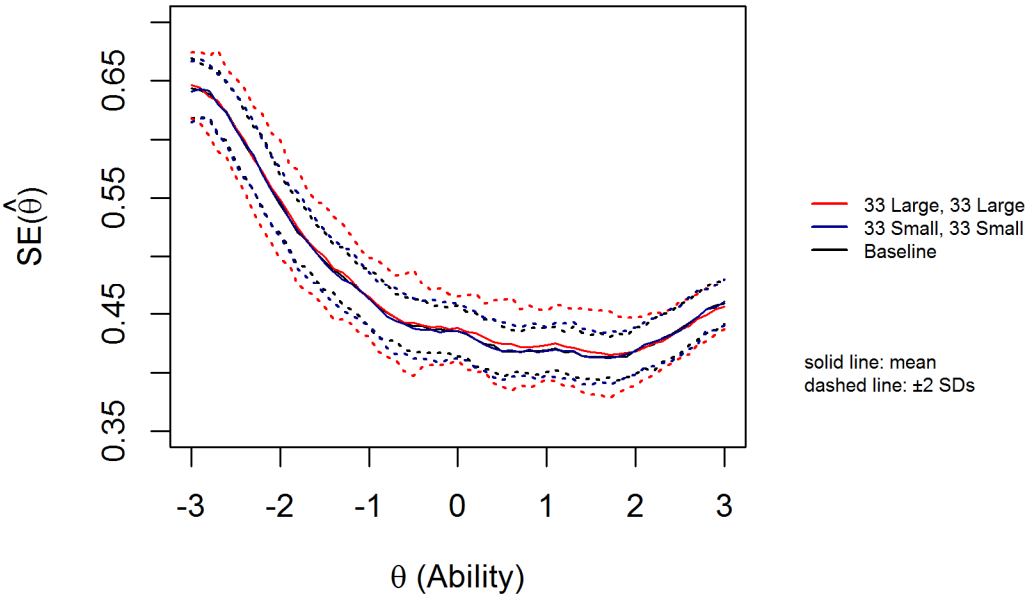


Figure B.2 Mean Standard Errors for 2 Conditions for Test Design 1-3-5

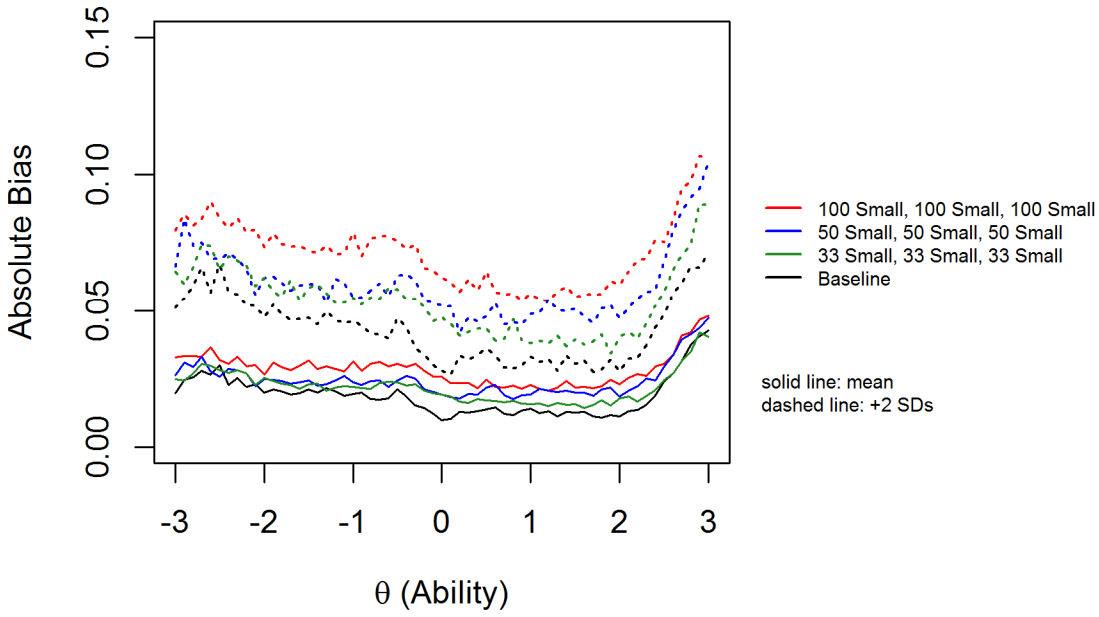


Figure B.3 Mean Absolute Bias for 3 Conditions for Test Design 1-3-5

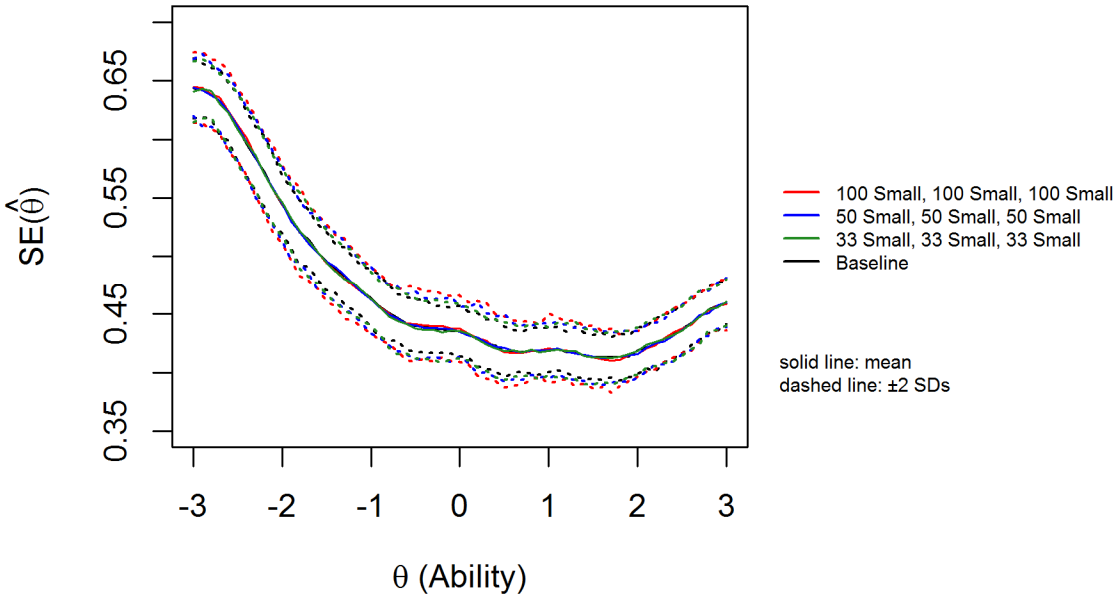


Figure B.4 Mean Standard Error for 3 Conditions for Test Design 1-3-5

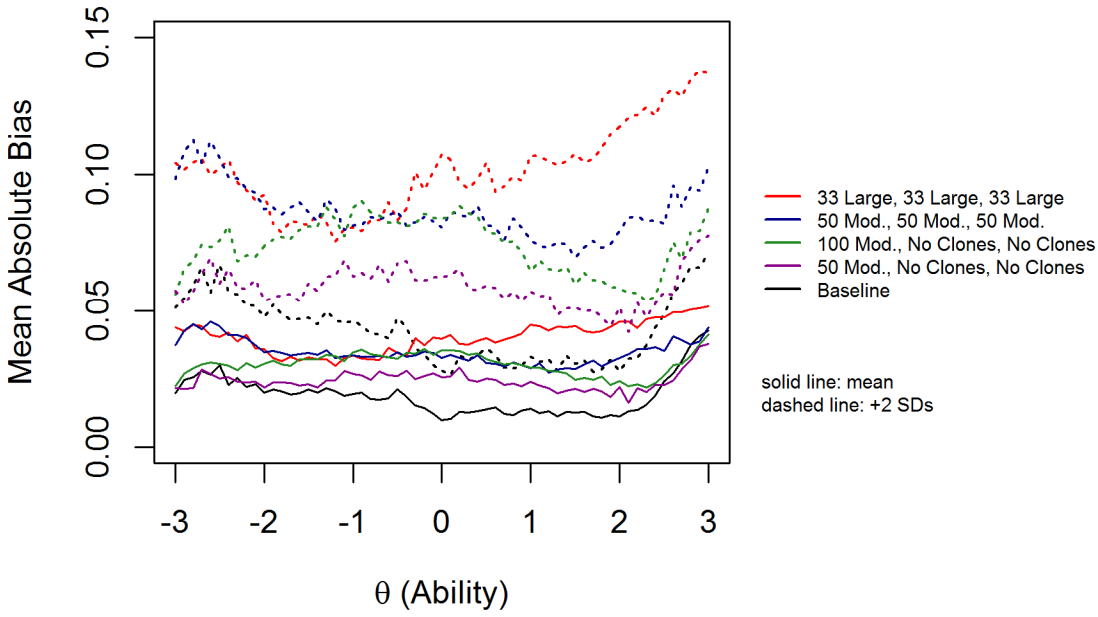


Figure B.5 Mean Absolute Bias for 4 Conditions for Test Design 1-3-5

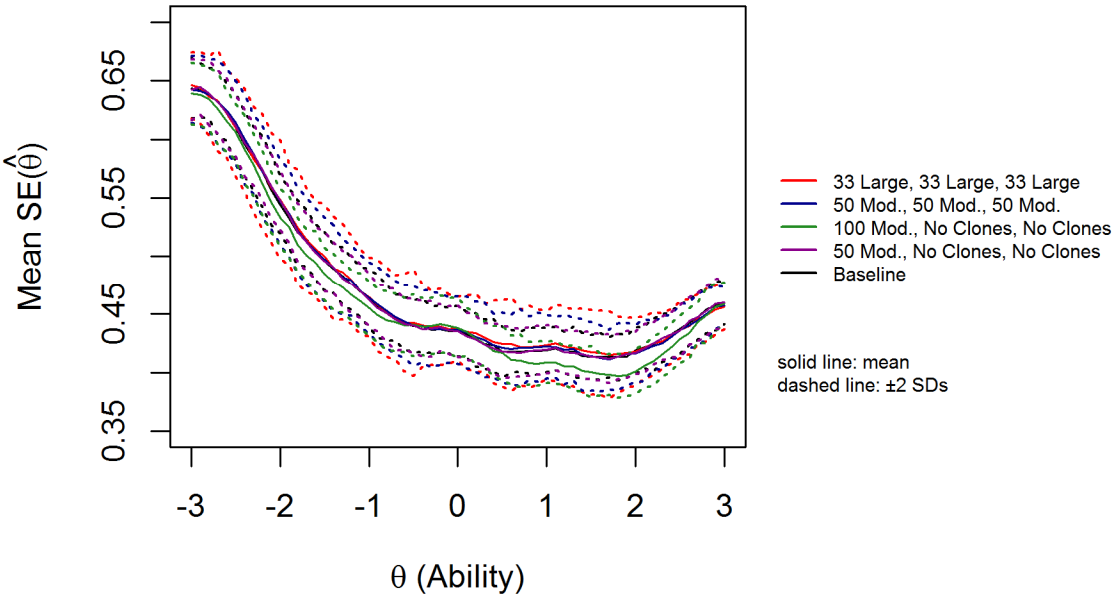


Figure B.6 Mean Standard Error for 4 Conditions for Test Design 1-3-5

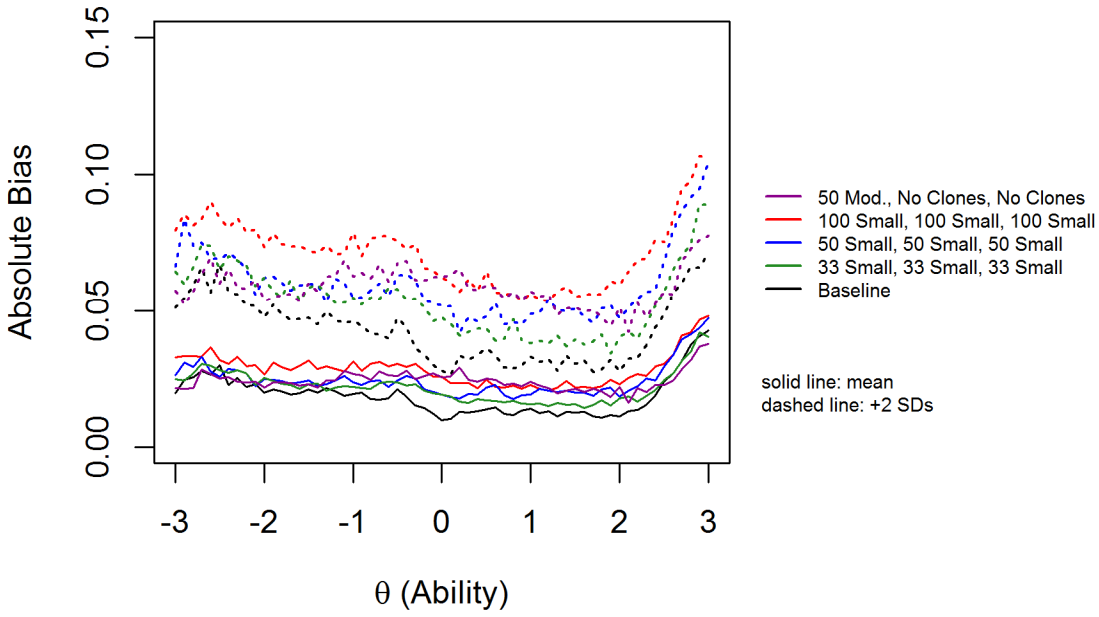


Figure B.7 Mean Absolute Bias for 4 Conditions for Test Design 1-3-5

APPENDIX C

RESULTS FOR TWO-STAGE TEST: DESIGN 1-2

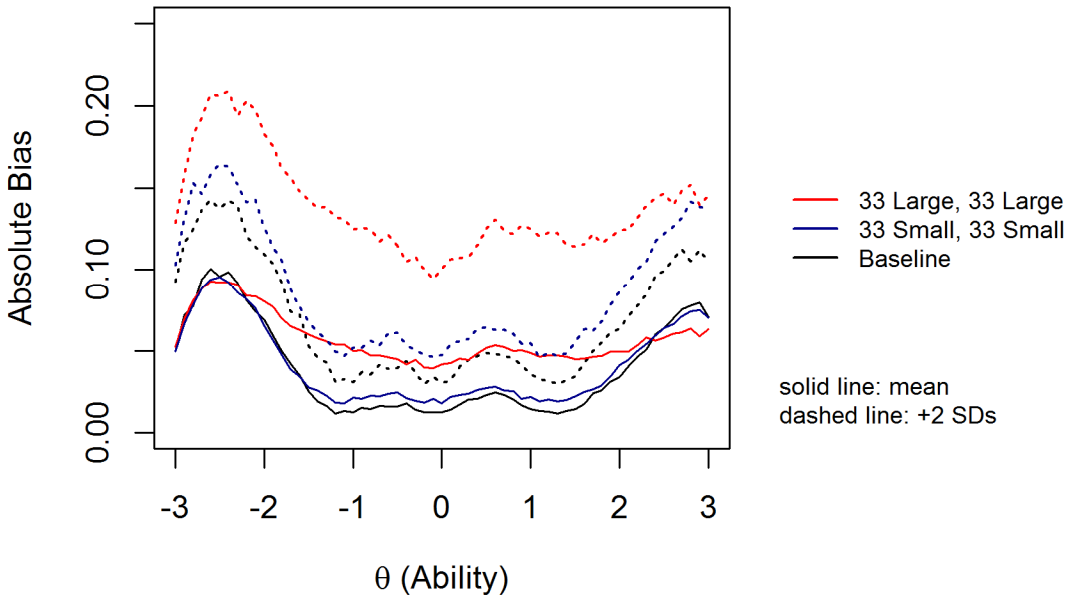


Figure C.1 Mean Absolute Bias for 2 Conditions for Test Design 1-2

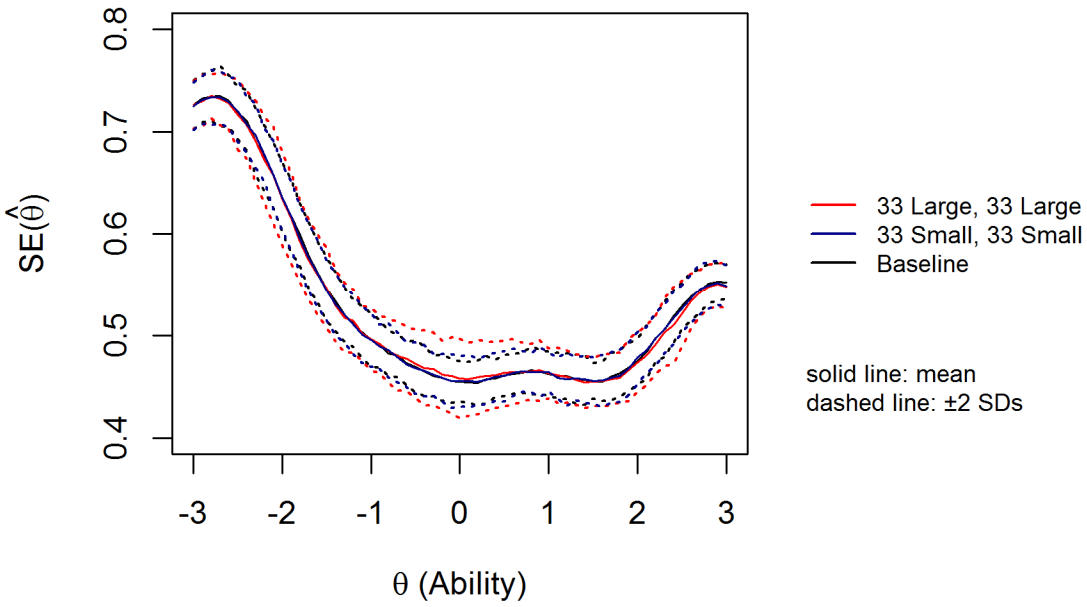


Figure C.2 Mean Standard Errors for 2 Conditions for Test Design 1-2

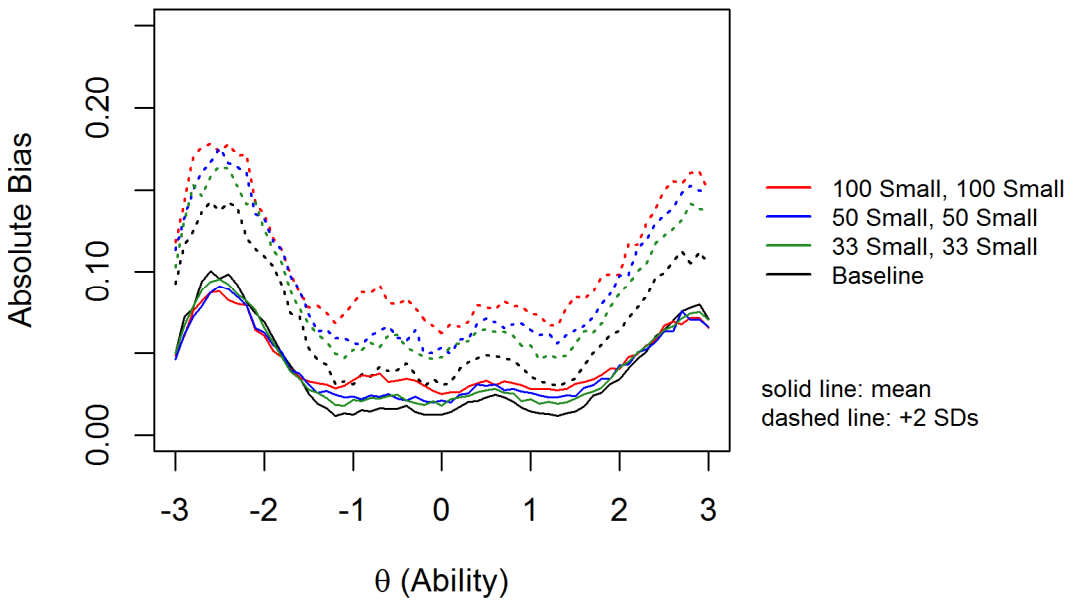


Figure C.3 Mean Absolute Bias for 3 Conditions for Test Design 1-2

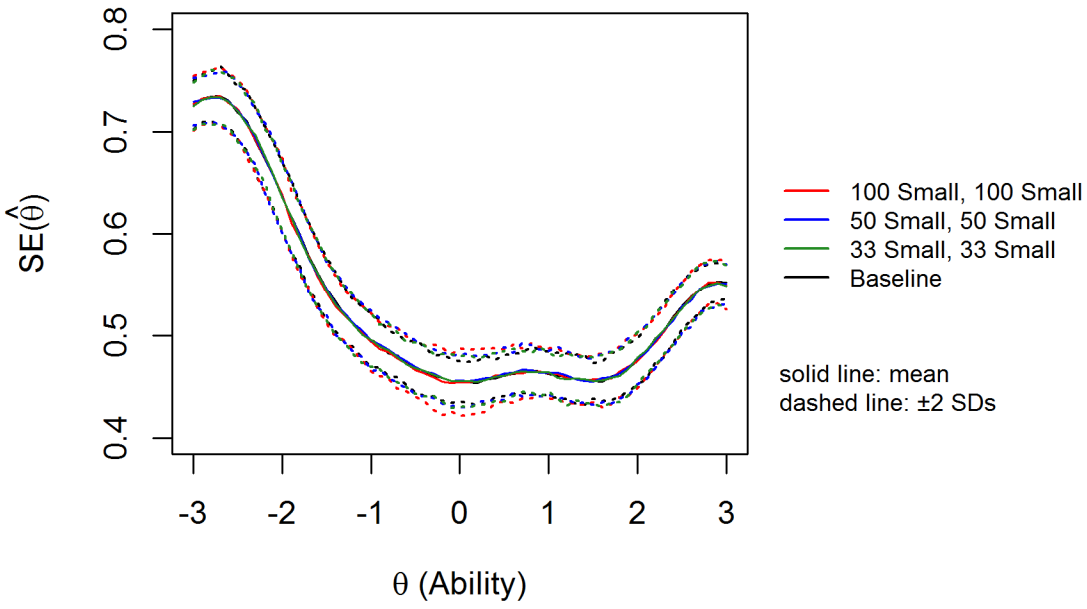


Figure C.4 Mean Standard Error for 3 Conditions for Test Design 1-2

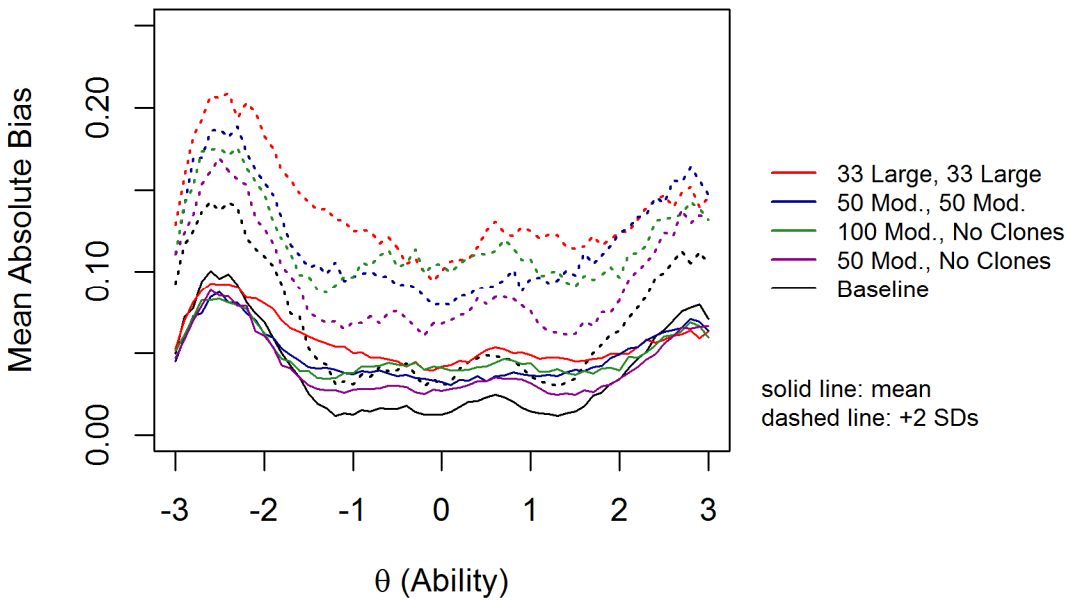


Figure C.5 Mean Absolute Bias for 4 Conditions for Test Design 1-2

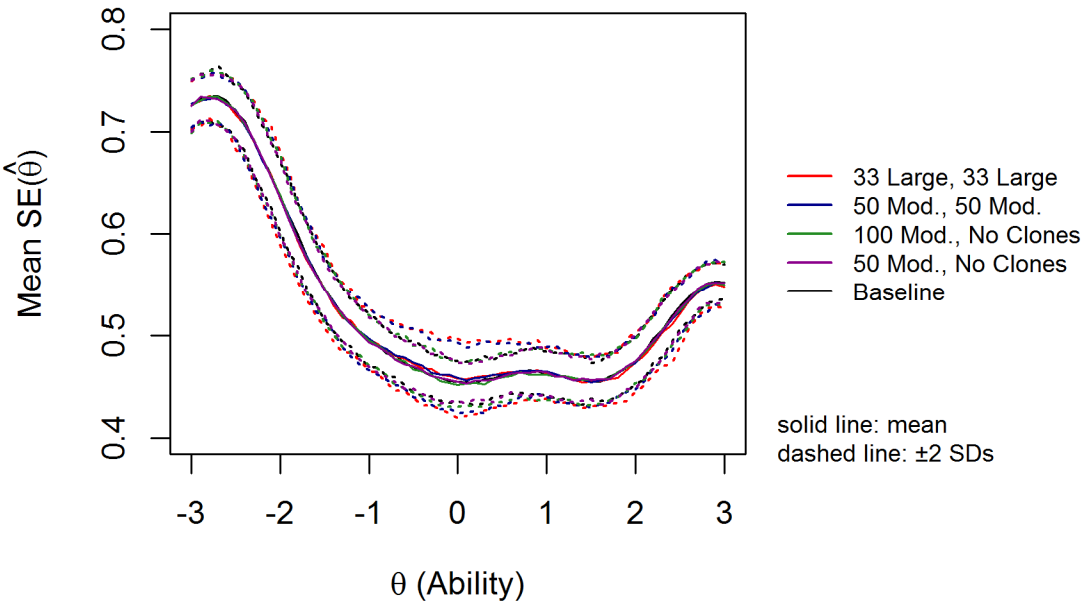


Figure C.6 Mean Standard Error for 4 Conditions for Test Design 1-2

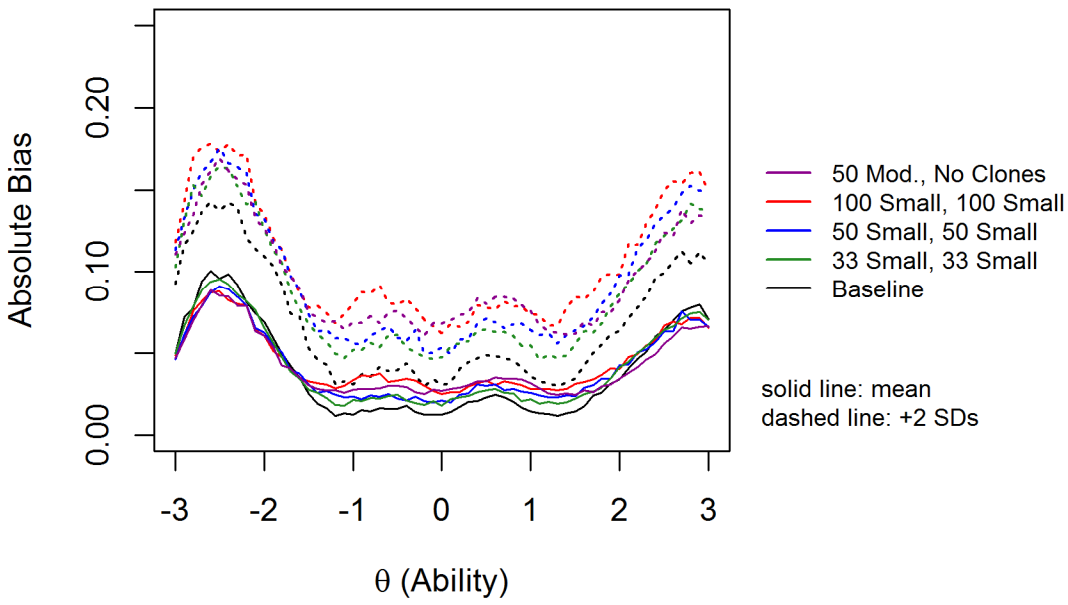


Figure C.7 Mean Absolute Bias for 4 Conditions for Test Design 1-2

APPENDIX D

RESULTS FOR TWO-STAGE TEST: DESIGN 1-2-4

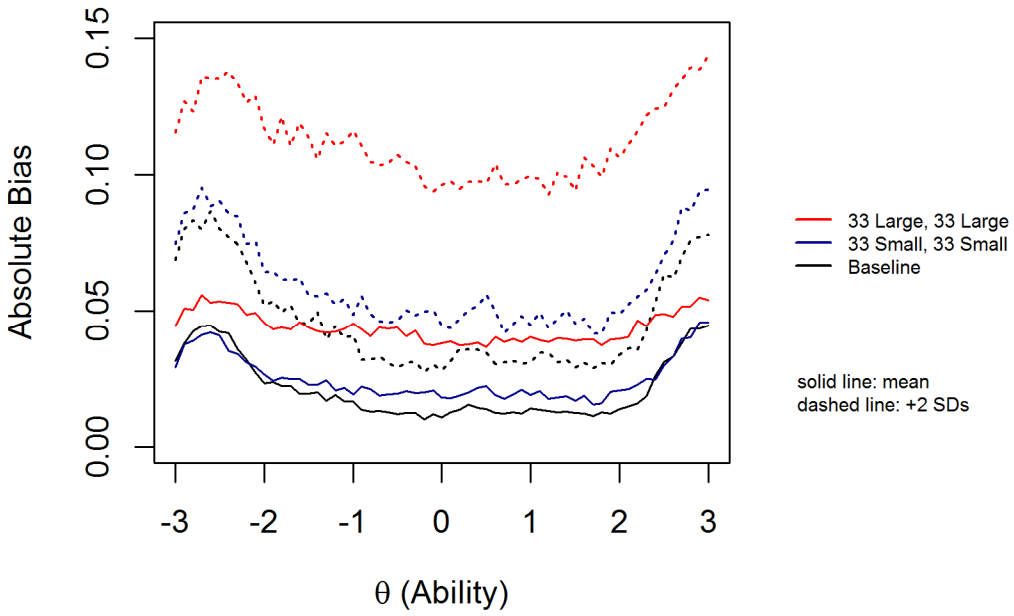


Figure D.1 Mean Absolute Bias for 2 Conditions for Test Design 1-2-4

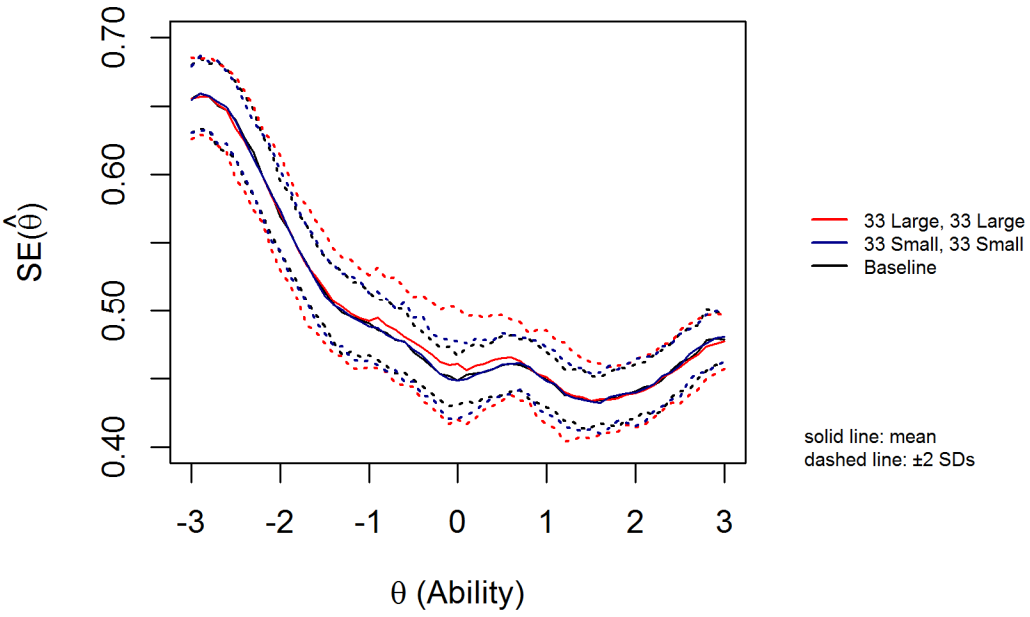


Figure D.2 Mean Standard Errors for 2 Conditions for Test Design 1-2-4

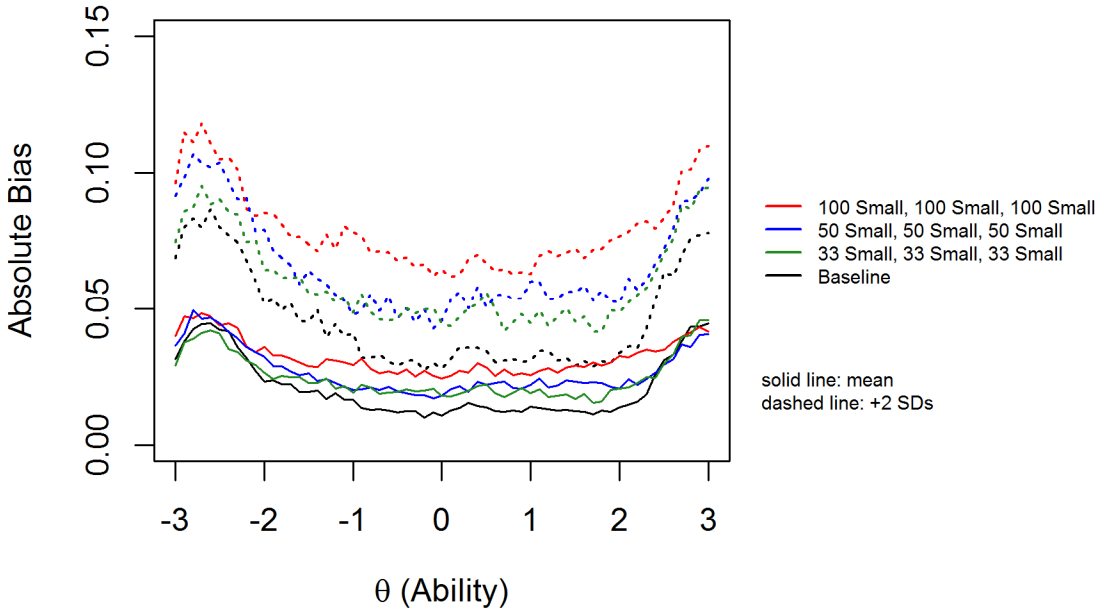


Figure D.3 Mean Absolute Bias for 3 Conditions for Test Design 1-2-4

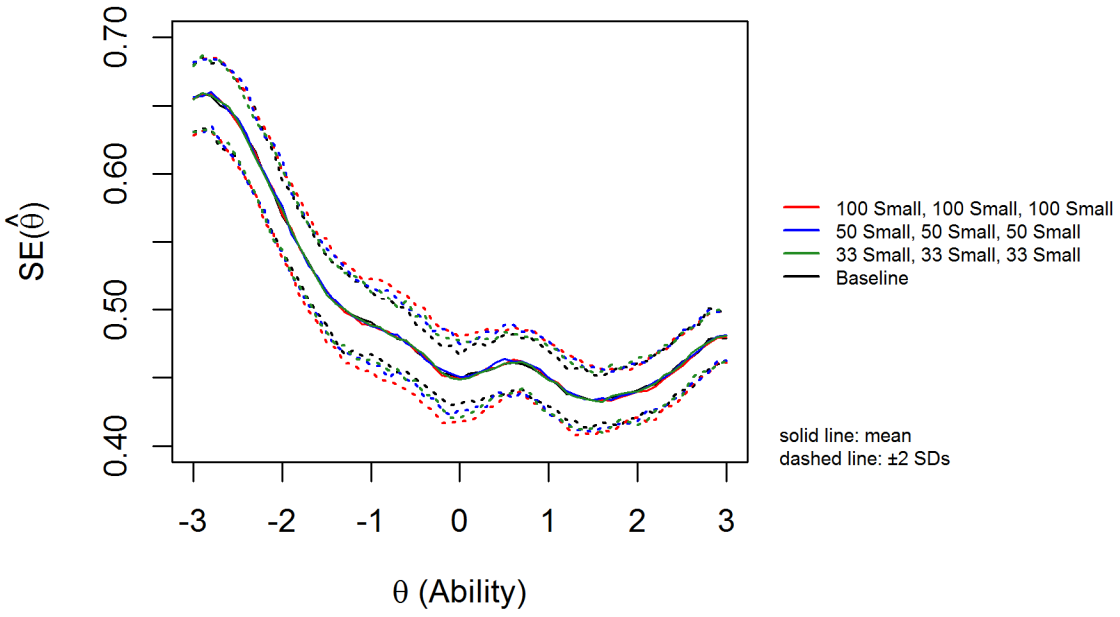


Figure D.4 Mean Standard Error for 3 Conditions for Test Design 1-2-4

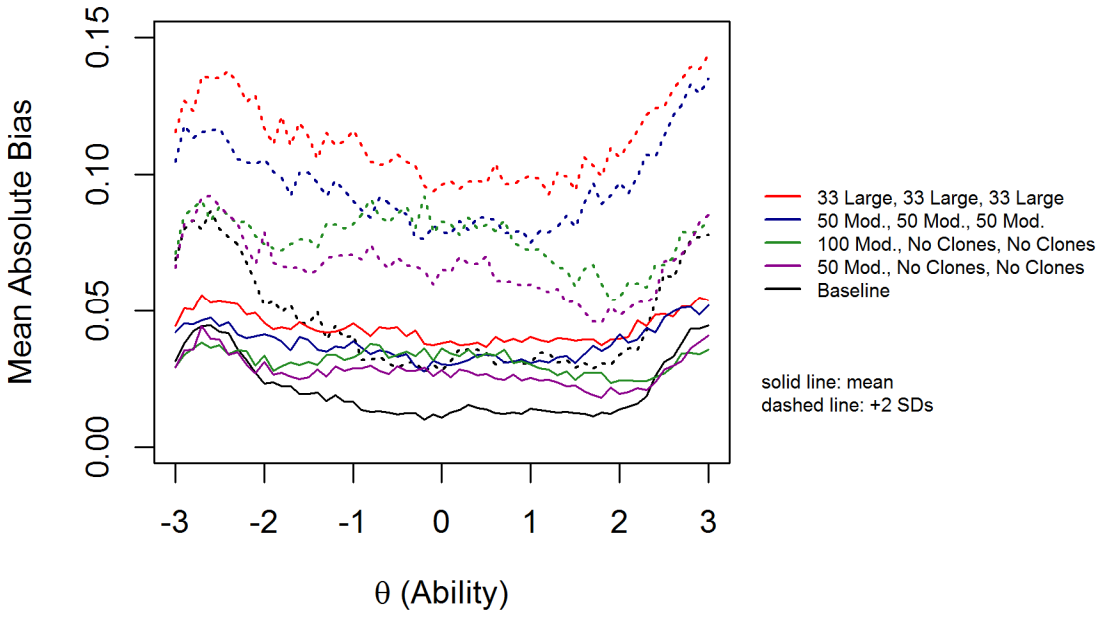


Figure D.5 Mean Absolute Bias for 4 Conditions for Test Design 1-2-4

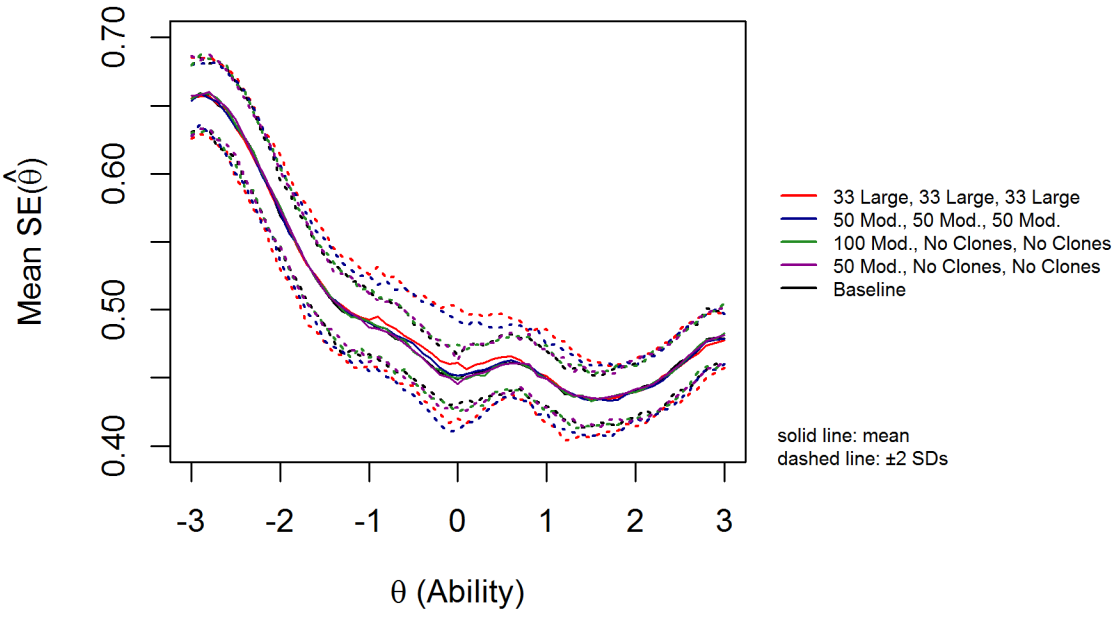


Figure D.6 Mean Standard Error for 4 Conditions for Test Design 1-2-4

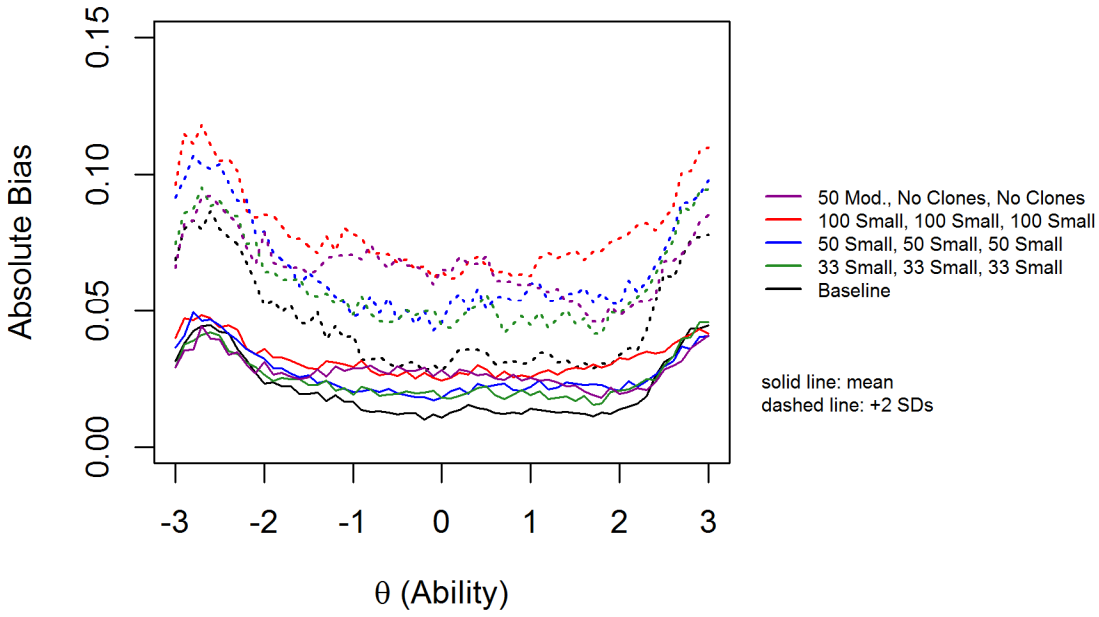


Figure D.7 Mean Absolute Bias for 4 Conditions for Test Design 1-2-4

REFERENCES

American Institute of Certified Public Accountants (n.d.). *How is the CPA exam scored?* Retrieved November 25, 2012, from http://www.aicpa.org/BECOMEACPA/CPAEXAM/PSYCHOMETRICSANDSCORING/Pages/PsychometricsandScoring.aspx

Arendasy, M., Sommer, M., & Ponocny, I. (2005). Psychometric approaches help resolve competing cognitive models: When less is more than it seems. *Cognition and Instruction, 23*(4), 503-521.

Bejar, I. I. (1993). A generative approach to psychological and educational measurement. In N. Frederiksen, R. J. Mislevy & I. I. Bejar (Eds.), *Test Theory for a New Generation of Tests* (1st ed., pp. 323-357). Hillsdale, NJ: Lawrence Erlbaum.

Bejar, I. I. (2010). Recent development and prospects in item generation. In S. E. Embretson (Ed.) , *Measuring Psychological Constructs: Advances in Model-Based Approaches* (pp. 201-226). Washington, DC US: American Psychological Association.

Bejar, I. I., Lawless, R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2003). A feasibility study of on-the-fly item generation in adaptive testing. *The Journal of Technology, Learning and Assessment, 2*(3).

Bennett, R. E. (2002). Inexorable and inevitable: The continuing story of technology and assessment. *Journal of Technology, Learning, and Assessment*, 1(1).

Bridgeman, B. (1998). Fairness in computer-based testing: What we know and what we need to know. In *New directions in assessment for higher education: Fairness, access, multiculturalism, & equity* (FAME) (pp. 4-11) (The GRE, FAME Report Series, Vol. 2). Princeton, NJ: Educational Testing Service.

Bridgeman, B., & Cline, F. (2004). Effects of differentially time-consuming tests on computer-adaptive test scores. *Journal of Educational Measurement, 41,* 2, 137-148.

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

Cronbach, L. J. & Gleser, G. C. (1965). *Psychological tests and personnel decisions.* Urbana: University of Illinois Press.

Daniel, R. C., & Embretson, S. E. (2010). Designing cognitive complexity in mathematical problem-solving items. *Applied Psychological Measurement, 34*(5), 348-364.

Deane, P., Graf, E. A., Higgins, D., Futagi, Y., & Lawless, R. (2006). *Model analysis and model creation: Capturing the task-model structure of quantitative domain items* (Research Report No. RR-06-01). Princeton, NJ: Educational Testing Service.

Drasgow, F., Luecht, R. M., & Bennett, R. (2006). Technology and testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 471-516). Washington, DC: American Council on Education.

DuBois, P. H. (1970). *A history of psychological testing*. Boston: Allyn & Bacon.

Educational Testing Service (n. d.). *GRE Revised General Test: Frequently asked questions*. Retrieved November, 25, 2012, from http://www.ets.org/gre/revised_general/faq/?viewfaq=faq6

Embretson (Whitely), S. (1983). Construct validity: construct representation versus nomothetic span. *Psychological Bulletin*, *93*, 179-197.

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika, 64*(4), 407-433.

Embretson, S. E., & Daniel, R. C. (2008). Understanding and quantifying cognitive complexity level in mathematical problem solving items. *Psychology Science Quarterly, 50*(3), 328-344.

Enright, M.K., Morley, M., & Sheehan, K.M. (2002). Items by design: The impact of systematic feature variation on item statistical characteristics. *Applied Measurement in Education, 15*(1), 49-74.

Enright, M. K., & Sheehan, K. M. (2002). Modeling the difficulty of quantitative reasoning items: Implications for item generation. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 129-158). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Fischer, G. H. (1973). Linear logistic test model as an instrument in educational research. *Acta psychologica, 37*, 359-374.

Glas, C. A.W. & van der Linden,W. J. (2001). *Modeling variability in item parameters in item models* (Research Report 01-11). Enschede: University of Twente.

Glas, C. W. & van der Linden, W. J. (2003). Computerized Adaptive Testing With Item Cloning. *Applied Psychological Measurement, 27*(4), 247.

Graduate Management Admission Council (n. d.) *GMAT: Test structure and overview*. Retrieved November 25, 2012, from http://www.mba.com/the-gmat/test-structure-and-overview.aspx?WT.svl=HPStructureandOverview

Green, B. F. (1983). The promise of tailored tests. In H. Wainer & S. Messick (Eds.), *Principals of Modern Psychological Measurement* (pp. 69-81). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

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

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

Jodoin, M. G., Zenisky, A. L., & Hambleton, R. K. (2006). Comparison of the psychometric properties of several computer-based test designs for credentialing exams with multiple purposes. *Applied Measurement in Education, 19,* 3, 203-220.

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

Lord, F. M. (1977). A broad-range tailored test of verbal ability. *Applied Psychological Measurement 1*, 95-100.

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

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

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

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

Mead, A. (2006). An introduction to multistage testing [Special Issue]. *Applied Measurement in Education, 19*, 185-260.

Mills, C. N., & Stocking, M. L. (1996). Practical Issues in Large-Scale Computerized Adaptive Testing. *Applied Measurement In Education*, *9*(4), 287.

Mislevy, R. J., Sheehan, K. M., & Wingersky, M. (1993). How to equate tests with little or no data. *Journal of Educational Measurement*, *30*, 55-78.

Nathan, N.J. & Petrosino, A. (2003). Expert blind spot among preservice teachers. *American Educational Research Journal, 40*(4), 905-928.

Newstead, S. E., Bradon, P., Handley, S. J., Dennis, I., & Evans, J. S. B. T. (2006). Predicting the difficulty of complex logical reasoning problems. *Thinking & Reasoning, 12*(1), 62-90.

Partnership for Assessment of Readiness for College and Careers (January 9, 2012). *Frequently asked questions: PARCC item development procurement and assessment development*. Retrieved November 25, 2012 from http://www.parcconline.org/sites/parcc/files/PARCC%20Item%20Development%20ITN%20FAQs%20-%20Updated%2001-09-12.pdf

Patsula, L. N. (1999). *A comparison of computerized-adaptive testing and multi-stage testing.* Unpublished doctoral dissertation, University of Massachusetts at Amherst.

Powers, D. E. (2001). Test anxiety and test performance: Comparing paper-based and computer-adaptive versions of the Graduate Record Examinations (GRE) General Test. *Journal of Educational Computing Research, 24,* 3, 249-73.

R Development Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Sebrechts, M. M., Enright, M., Bennett, R.E., & Martin, K. (1996). Using algebra word problems to assess quantitative proficiency: attributes, strategies, and errors. *Cognition and Instruction*, 14(3), 285-343.

Sinharay, S., & Johnson, M. S. (2008). Use of item models in a large-scale admissions test: A case study. *International Journal of Testing, 8*(3), 209-236.

Smarter Balanced Assessment Consortium (n. d.) *Computer adaptive testing.* Retrieved November 25, 2012, from http://www.smarterbalanced.org/wordpress/wp-content/uploads/2011/12/Smarter-Balanced-CAT.pdf

Stocking, M. L., & Lewis, C. (1998). Controlling item exposure conditional on proficiency in computerized adaptive testing. *Journal of Educational Statistics, 23,* 1, 57-75.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: *Journal of Educational Measurement, 26*, 247-260.

U.S. Department of Education, National Center for Education Statistics. (2010). *Teachers' use of educational technology in U.S. public schools: 2009* (NCES 2010-040).

van der Linden, W. J. (2008). Some new developments in adaptive testing technology. *Journal of Psychology*, *216*, 3-11.

van der Linden, W. J. & Pashley, P. J. (2010). Item selection and proficiency estimation in adaptive testing. In W. J. van der Linden & C. A. W. Glas (eds.), *Elements of adaptive testing* (pp. 3-30). New York: Springer.

van der Linden, W. J. & Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics, 29,* 3, 273-291.

Vispoel, W. P., Rocklin, T. R., & Wang, T. (1994). Individual differences and test administration procedures: A comparison of fixed-item, computerized-adaptive, and self-adapted testing. *Applied Measurement in Education, 7,* 53-79.

Wainer, H. (2002). *On the automatic generation of test items.* In S. H. Irvine & P. C. Kyllonen (Eds.), Item generation for test development (pp. 287-305). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (1990). *Computerized adaptive testing: A primer*. Hillsdale, NJ: L. Erlbaum Associates.

Wainer, H. & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 3, 185-201.

Zenisky, A., Hambleton, R. K., & Luecht, R. M. (2010). Multistage testing: Issues, Designs, and Research. In W. J. van der Linden & C. A. W. Glas (eds.), *Elements of adaptive testing* (pp. 355-372). New York: Springer.