Evaluating the Performance of CATSIB in a Multi-Stage Adaptive Testing Environment

Mark J. Gierl
Hollis Lai
Johnson Li

Centre for Research in Applied Measurement and Evaluation
University of Alberta

FINAL REPORT

Submitted to:
Dr. Krista Breithaupt
Director, Research and Development
Medical Council of Canada

May 18, 2011

ABSTRACT

The purpose of this study is to evaluate the performance of CATSIB for detecting DIF when items in the matching and studied subtest are administered adaptively in the context of a realistic multi-stage adaptive test (MST). MST was simulated using a four-item module in a seven-panel administration, where the first panel contained a single four-item module that was common to all examinees. The second to seventh panels each contained three modules with four items per module at three different difficulty levels. Each examinee wrote seven modules thereby completing 28 items. Three independent variables, expected to affect DIF detection rates, were manipulated: item difficulty (easy, medium, hard modules), sample size (small—100-175 examinees per group; moderate—200-300 examinees per group; large—300-450 examinees per group), and balanced/unbalanced design (same or different sample sizes in each group). Two types of dependent variables were used to evaluate the performance of CATSIB, Type I error and power. CATSIB met the acceptable criteria, meaning that the Type I error and power rates met 5% and 80%, respectively, for the large reference/moderate focal sample and the large reference/large focal sample conditions. These results indicate that CATSIB can be used to consistently and accurately detect DIF on an MST, but only with moderate to large samples. Directions for future research on DIF in CAT are also discussed.

INTRODUCTION

Randy Bennett (2001) claimed, a decade ago, that no topic would become more central to innovation and future practice in large-scale assessment than computers and the internet. His prediction has proven to be accurate. Large-scale assessment and computer technology have evolved at a staggering pace since 2001. As a result many large-scale assessments, which were once administered in a paper-and-pencil format, are now administered by computer using the internet. Education Week's 2009 *Technology Counts*, for example, reported that almost half of the US states administered internet-based computerized educational assessments. Many popular and well-known assessments such as the Graduate Management Achievement Test (GMAT), the Graduate Record Exam (GRE), and the Test of English as a Foreign Language (TOEFL iBT), to cite but a few examples, are administered by computer over the internet.

Internet-based computerized assessment offers many advantages to examinees compared with more traditional paper-based assessments. For instance, computers support the development of innovative item types and alternative item formats (Sireci & Zenisky, 2006); items on computer-based tests can be scored immediately thereby providing examinees with instant feedback (Drasgow & Mattern, 2006); computers permit continuous and on-demand testing for examinees (van der Linden & Glas, 2010). But possibly the most important advantage of computer-based assessment is that it allows examiners to evaluate more complex performances by integrating test items and digital media to substantially increase the types of knowledge, skills, and competencies that can be measured (Bartram, 2006; Breithaupt, Mills, & Melican, 2006; Zenisky & Sireci, 2002).

The advent of computer-based testing has also raised new challenges, particularly in the area of test fairness. Fairness is a broad and encompassing topic of importance and consequence in educational and psychological testing. In the 1999 *Standards for Educational and Psychological Testing*, four characterizations or aspects of test fairness are presented. First, fair tests must be free

from bias. Bias occurs when tests yield scores or promote score interpretations that result in different meanings for members of different groups. Second, test fairness requires that examinees receive just and equal treatment in the testing process. To achieve this outcome, both the test and the testing context must be considered when scores are interpreted for examinees. Third, test fairness requires equity in the outcomes of testing. Examinees must be given the chance to demonstrate their proficiency on the construct the test is designed to measure. Fourth, test fairness implies that examinees have had the opportunity to learn the content covered on the exam.

Differential item functioning (DIF) analyses can yield information about bias, which is the first characterization of fairness cited in the *Standards* (1999). The development and application of DIF detection methods reflect, in large part, a response to the legal and ethical need to assess examinees without bias. To conduct DIF analyses, examinees are first divided into two groups, a *reference* and *focal* group. DIF analysis then involves administering a test, matching members of the reference and focal group on a measure of ability derived from that test, and using statistical procedures to identify items that function differentially between the two groups. An item exhibits DIF when examinees from the reference and focal groups differ in the probability of answering that item correctly, after controlling for the measure of ability derived from the test.

It is particularly important to conduct DIF analyses with computer adaptive tests (CAT). The importance stems from the aforementioned legal and ethical need to evaluate examinees without bias. This need is simply amplified for internet-based CAT because these tests, increasingly, are being administered to large numbers of examinees around the world. Because testing has become a global enterprise, heterogeneous samples of examinees with different languages, cultures, educational backgrounds, learning opportunities, knowledge, skills, and access to computers and technology are writing the same exams that are expected to yield the same score interpretations. For example, Phillippe Grosskost, Managing Director of ETS Global for Europe, the Middle East, and Africa, claimed

that “Scores on the TOEIC and TOEFL tests mean exactly the same thing regardless of whether the test was taken in Indonesia, Argentina, Hungary, or Egypt” (Educational Testing Service, 2007). This strong assertion highlights the importance of testing without bias. The need for DIF-free item administration also stems from the adaptive nature of CAT. As Zwick (2000, 2010) noted, examinees write fewer items in an adaptive testing context, meaning that each item contributes more to the final ability estimate. The presence of item bias, therefore, could exert a stronger affect on the examinees’ estimates of ability. Bias could also affect the order of item administration, given that the selection of items on an adaptive test is determined, in part, by the examinees’ response to the previous item.

It is also particularly challenging to conduct DIF analyses with CAT. CAT requires large numbers of items because banks are needed to permit continuous testing while, at the same time, minimizing item exposure. As a result, these large item banks must first be developed and then continually replenished to minimize item exposure and maintain test security while allowing for continuous test administration. At the same time, policies, procedures, and reviews must be implemented to ensure each item meets the basic standards associated with fairness and equity. Sensitivity reviews—which entails having panels of content specialists review each item—are conducted to ensure items used for CAT meet these basic standards (Educational Testing Service, 2009; Zieky, 1993). Sensitivity reviews are also informed by the outcomes from DIF analyses. Unfortunately, the number of examinees who write any one item on an adaptive test may be small, particularly when the item bank is large, relative to the items on a paper-based exam. As a result, DIF methods designed to help monitor fairness must function in diverse testing environments and, often, when the total number of items in the bank is large but the number of examinees who respond to any one of those items is relatively small.

Given that DIF detection is both an important and challenging undertaking with CAT, it comes as *some surprise* that little research has been conducted on this topic in the last decade. Or, said differently, research on DIF in CAT is not thriving. In 2000, Zwick published a seminal chapter on DIF in

CAT as part of the book *Computer Adaptive Testing: Theory and Practice* (van der Linden & Glas, 2000). She reviewed the three DIF detection methods that, at the time, were considered the main CAT DIF procedures—the Zwick, Thayer, and Wingersky CAT DIF method (ZTW), the CAT version of the empirical Bayes Mantel-Haenszel DIF method of Zwick, Thayer, and Lewis (ZTL), and CAT for SIBTEST by Nandakumar and Roussos (CATSIB). Zwick provided a review of each method in its original, non-adaptive version. Then, she described each method in its modified, adaptive version. Finally, she presented some empirical results from simulation studies to support each CAT DIF method.

A decade later, in 2010, *Computer Adaptive Testing: Theory and Practice* (van der Linden & Glas, 2000) was revised and updated, and published as *Elements of Adaptive Testing* (van der Linden & Glas, 2010). The new volume featured revised chapters from many of the original authors as well as some new chapters. Zwick's chapter on DIF in CAT was included in the 2010 volume. The most striking feature of Zwick's revised chapter was how *little* the area of DIF in CAT had changed over the last 10 years. In the decade since the publication of her first chapter, no new DIF methods for CAT were introduced in Zwick's review. Moreover, only 10 new references were included in her updated manuscript (out of a total of 57 references in the 2010 chapter), of which five were published prior to the publication of Zwick's first chapter in 2000, five were published after 2000, and only two of the five references published after 2000 were found in referred journals (the other three citations appeared in technical reports). In short, relatively little research has been conducted on DIF in CAT since 2000 despite the explosion of research on and application of computer-based and computer adaptive testing over the same time period.

Thus, the purpose of the current study is to begin expanding the research on DIF in CAT by evaluating the performance of CATSIB in a multi-stage adaptive testing environment. CATSIB (Nakadumar & Roussos, 2001, 2004), a modification of SIBTEST (Shealy & Stout, 1993) intended for CAT, is a statistical procedure used to first match reference and focal group examinees on regression-

corrected IRT-based ability estimate, and then compare the examinees on a weighted mean difference to determine the presence of DIF. Matching on ability is particularly important for DIF detection because it ensures that only examinees who achieve the same ability estimates are compared to one another in the analysis. CATSIB was selected for two reasons. First, CATSIB, which is one of Zwick's (2000, 2010) three main CAT DIF methods, has received limited empirical evaluation in a small number of CAT environments. To-date, CATSIB has only been evaluated in an item pretesting context in two studies (see Nandakumar & Roussos, 2001, 2004; Lei, Chen, & Yu, 2006). Nandakumar and Roussos first studied the performance of CATSIB on pretest DIF items administered in a non-adaptive manner. In other words, adaptation affected the non-DIF items used to match examinees, but not the pretest DIF items themselves. Hence, they evaluated the performance of CATSIB for DIF detection in a testing environment where the matching subtest (i.e., the subtest containing the non-DIF items) was created adaptively and assumed to be free from DIF whereas the studied subtest (i.e., the subtest containing the DIF items) was not created adaptively and assumed to contain DIF. This design can be used when items on the matching subtest are either known or assumed to be free from DIF. The addition of new, presumably non-counting items that do not contribute to the examinees' ability estimate, are then evaluated for DIF. In this context, CATSIB performed well across a variety of study conditions in the Nandakumar and Roussos study. When DIF was large, power was relatively high and Type I error was relatively low. When DIF was moderate, power decreased and Type I error increased.

Lei et al. (2006) also used a pretest design to compare the Type I error and power rates of CATSIB, logistic regression, and the IRT likelihood-ratio test for identifying unidirectional and non-unidirectional DIF in a CAT environment. As with Nandakumar and Roussos, Lei et al. evaluated the performance of CATSIB in a testing environment where the matching subtest was created adaptively and assumed to be free from DIF and a studied subtest that was not created adaptively. Again, CATSIB performed well across those conditions that were similar to Nandakumar and Roussos for detecting

unidirectional (i.e, uniform) DIF but CATSIB was less effective when attempting to identify non-unidirectional (i.e., non-uinform) DIF. Lei et al. also found that CATSIB was more accurate for detecting unidirectional DIF when the samples sizes were balanced (i.e., 500 examinees in the reference and focal groups) compared to conditions where the sample sizes were dramatically unbalanced (i.e., 100 focal group examinees; 900 reference group examinees). These two studies provide important information on key variables that affect CATSIB DIF detection in a pretesting context using adaptively administered non-DIF items and non-adaptively administered DIF items. Unfortunately, their results may not generalize to the performance of CATSIB DIF detection in a purely adaptive context, meaning when items in both the matching and studied subtests are administered adaptively. Therefore, additional research is required to evaluate the performance of CATSIB in more diverse CAT environments, particularly when both DIF and non-DIF items are administered adaptively.

The second reason CATSIB was selected is that it is a psychometric method embedded in the multidimensional differential item functioning analysis paradigm proposed by Roussos and Stout (1996a) that unifies substantive and statistical DIF analyses by linking both to the Shealy-Stout multidimensional model for DIF (Shealy & Stout, 1993). This paradigm is unique in the DIF research area because it is designed to help researchers and practitioners understand *why* DIF occurs. Typically, DIF statistical analyses are followed by sensitivity reviews to identify the sources and probable causes of DIF. Reviewers are asked to study DIF items and describe why these items are more difficult for one group of examinees compared to another (see, for example, Camilli & Shepard, 1994, p. xiii). But researchers found that reviewers were generally poor at predicting which items would function differently across groups (e.g., Englehard, Hansche, & Rutledge, 1990; Plake, 1980). Practitioners also claimed it is difficult to interpret DIF using the judgmental approach (e.g., Camilli & Shepard, 1994; O'Neill & McPeek, 1993; *Standards for Educational and Psychological Testing*, 1999). From these well-documented outcomes, Roussos and Stout (1996a) concluded: “Attempts at understanding the

underlying causes of DIF using substantive analyses of statistically identified DIF items have, with few exceptions, met with overwhelming failure" (p. 360). To overcome this problem, they proposed the multidimensional DIF analysis paradigm. It is a confirmatory approach conducted in two stages. The first stage is a substantive analysis where DIF hypotheses are generated. The second stage is a statistical analysis of the DIF hypotheses. By combining substantive and statistical analyses in a confirmatory framework, researchers can identify and study the sources and probable causes of DIF (e.g., Gierl, 2005; Gierl, Bisanz, Bisanz, Boughton, & Khaliq, 2001; Gierl & Khaliq, 2001; Gierl, Bisanz, Bisanz, & Boughton, 2003). That is, DIF analyses are ultimately conducted to help researchers and practitioners understand why items elicit group difference and, hence, when bias is occurring. The DIF analysis paradigm serves as one tangible framework for pursuing the question of why DIF occurs. But before researchers and practitioners address the question of why DIF occurs, they must be able to consistently and accurately identify items that produce these group differences. The results from CATSIB can be placed in the DIF analysis paradigm to understand why DIF occurs in CAT. To-date, however, no research has been conducted with CATSIB to evaluate its performance for identifying DIF when items from both the matching and studied subtests are administered adaptively.

Therefore, the purpose of the current study is to evaluate the performance of CATSIB for detecting DIF when both the matching and studied subtest items are administered adaptively in the context of a realistic multi-stage adaptive testing environment. Our research contributes to the study of DIF in CAT in two ways. We will evaluate CATSIB when items on the studied and matching subtests are administered adaptively. Nandakumar and Roussos (2001, 2004) and Lei et al. (2006) only evaluated CATSIB when items on the matching subtest were administered adaptively. We will also evaluate CATSIB in the realistic and, increasingly, popular CAT context of multi-stage adaptive testing (MST). MST involves adaptive selection of sets of items, rather than single items, for sequential administration. Just as DIF on CAT can adversely affect the adaptive item administration procedures,

the consequences of DIF on MST could exert a negative affect on the examinees' estimate of ability because bias could affect the order of administration for the sets of items on an MST. To-date, however, there has been no research on DIF in MST (Richard Luecht, personal communication, December 13, 2010).

CATSIB AND MULTI-STAGE ADAPTIVE TESTING

Overview of CATSIB

CATSIB is used to test the statistical hypotheses H_o: $\hat{\beta}_{UNI} = 0$ versus H_1: $\hat{\beta}_{UNI} \neq 0$, where $\hat{\beta}_{UNI}$ is the parameter specifying the amount of DIF that occurs for an item when examinees in the reference and focal group are compared. More specifically, $\hat{\beta}_{UNI}$ serves as a measure of the expected probabilistic difference of a correct response between examinees in the reference and focal group. That is, $\hat{\beta}_{UNI} = ES_R(\theta) - ES_F(\theta)$, where $ES_R(\theta)$ and $ES_F(\theta)$ are the expected scores in the reference and focal groups, respectively, conditional on the matching subtest. But because $ES_R(\theta)$ and $ES_F(\theta)$ contain bias due to measurement error, a regression correction procedure is used to adjust these expected item scores so they yield more reliable estimates. The adjusted expected scores are denotes as $ES_R\ (\theta^*)$ and $ES_F\ (\theta^*)$ for the reference and focal groups, respectively. The adjusted scores more accurately reflect examinees of equal ability levels across groups and, thus, are more meaningful for comparing group differences on the studied items. CATSIB then uses the weighted average difference of these adjusted scores (weighted by the proportion of examinees obtaining matching subtest score θ^*) to estimate the DIF index $\hat{\beta}_{UNI}$, where $\hat{\beta}_{UNI} = \sum_{\theta^*=\theta^*_{Min}}^{\theta^*_{Max}} \left[\left(ES_R\ (\theta^*) - ES_F\ (\theta^*)\right)\frac{N_R(\theta^*)-N_F(\theta^*)}{N}\right]$. In this formula, θ^* is the examinees' regression-corrected ability estimate on the matching subtest, $N_R(\theta^*)$ and $N_F(\theta^*)$ are the number of examinees obtaining matching subtest score θ^* from the reference and focal groups, respectively, and N is the total number of examinees. The $\hat{\beta}_{UNI}$ index is

distributed as approximately normal assuming a null hypothesis of no DIF. The standard error of $\hat{\beta}_{UNI}$ is given as $\hat{\sigma}_{\hat{\beta}UNI} = \left[\sum_{\theta^*=\theta^*_{Min}}^{\theta^*_{Max}} \left(\frac{N_R(\theta^*)-N_F(\theta^*)}{N}\right)^2 \left(\frac{\widehat{\sigma^2_R}(\theta^*)}{N_{R(\theta^*)}} + \frac{\widehat{\sigma^2_F}(\theta^*)}{N_{F(\theta^*)}}\right)\right]^{1/2}$.

Overview of Multi-Stage Adaptive Testing

Multi-stage adaptive testing (MST) involves adaptive selection on sets of items, rather than a single item as with CAT, for sequential administration. It is a form of adaptive testing that has been evaluated, scrutinized, and empirically supported (see, for example, the studies presented in *Applied Measurement in Education*, Special Issue: Multi-stage Testing, 2006). As a result, MST is growing in popularity and, currently, is implemented in well-known testing programs such as the American Institute of Certified Public Account's Uniform CPA examination, which is used to license public accountants in the United States, and the Medical Council of Canada's Qualifying Exam Part I, which is used to admit medical students into supervised clinical practice in Canada. MST has many of the benefits of CAT, including reducing test length, improving test administration standardization, increasing measurement precision, improving testing security, and allowing for testing on-demand, with the added benefit of control in the assembly stage of test development by permitting content specialists to create multiple test forms that can be reviewed prior to test administration (Luecht & Nungester, 1998; Zenisky, Hambleton, & Luecht, 2010).

Many different versions and variations of MST are permissible. However, some key concepts do exist. A block or set of items is referred to as a *module* (Luecht & Nungester, 1998). Each module contains a set of items that adhere to specific content requirements while also meeting strict statistical specifications. These modules then become part of a computer-adaptive process, in which modules with different difficulty levels are administered to examinees. Although each examinee completes all of the items administered within a module, any two examinees need not receive the same module or sequence of modules because the module and its administration order is based on each examinees'

ability estimate. The modules are administered in stages called *panels* to facilitate the adaptive process. In each panel, the module is created to meet a specific level of difficulty, where the difficulty level is matched to the examinee's provisional ability level as estimated from their performance on the modules administered during the previous panel. Within any one panel, the modules typically have two of more difficulty levels and, hence, permit adaptive sequencing for test administration. After the examinee completes the items for one module, the ability estimate is updated and, based on the estimate, the module in the next panel that provides the most measurement information is presented to the examinee.

An example of a three-panel, seven-module adaptive MST is presented in Figure 1. In this example, examinees are all administered the same starter module in panel 1 and then move to the next module in panel 2 according to their current ability estimate. Figure 1 illustrates how the modules are linked together by three primary pathways examinees may take, depending on their current ability estimate. For example, if an examinee does poorly on module A, then she or he will be administered module B in panel 2. If the examinee then does well on module B, then she or he will be administered module F in panel 3. The modules from left to right cover the same content areas and only differ in difficulty level.

METHODS

We evaluated the performance of CATSIB for detecting DIF when both the matching and studied subtest items are administered adaptively in the context of a simulated MST. We begin by describing the data generation process and the independent variables (module difficulty level; sample size; balanced/unbalanced design) and then we describe the dependent variables (Type I error; power) used in our study to evaluate the performance of CATSIB.

Data Generation and Independent Variables

For DIF studies using paper-based test administration (i.e., studies with no adaptation, where every examinees writes every item), the magnitude of DIF is controlled during the simulation of sample

responses, meaning the examinee responses vectors can be modified to fit the expected magnitude of DIF. This approach helps ensure that an expected level of DIF is generated for any given item in the study. For DIF studies using CAT administration, the DIF items cannot be generated in this manner because the examinee item responses are dynamic thereby affecting the expected level of DIF. That is, the simulated examinee item responses produced with an adaptive testing process affect the magnitude of DIF. To address this problem, the DIF items were generated according to a normal distribution, with the mean of $\hat{\beta}_{UNI}$ = 0.11 and standard deviation $\hat{\beta}_{UNI}$ = 0.04. This DIF magnitude is considered large (Nandakumar & Roussos, 2001, 2004), indicating that matched reference and focal group examinees differ, on average, by 1/9 of a score point. Only items that produced large DIF were simulated and hence evaluated in our study.

A MST environment was developed using the R programming language (R Development Core Team, 2011). Our simulated MST used a four-item module in a seven-panel administration, where the first panel contained a single module that was common to all examinees while the second through seventh panels each contained three modules with items at three different difficulty levels (see Figure 2). Using this structure, each examinee completed seven modules and wrote a total of 28 items. Each module contained items at three levels of difficulty—easy, medium, and hard. Because item difficulty could affect CATSIB DIF detection, particularly when variation among the levels differs, it served as the first independent variable in this study. The parameter estimates for the generated items at each of the three levels is provided in Table 1. The items were generated using a 2PL IRT model. The a-parameter was generated with a mean of 0.80 (SD=0.20) across the three difficulty levels indicating that, overall, the quality of the items was high, but that the bank also contained items with a range of discrimination power as one might expect in an operational computer-based testing situation. The b-parameter was variable across three levels of difficulty: The easy level had a mean of -1.25 (SD=0.50); the medium level had a mean of -0.25 (SD=0.25); the hard level had a mean of 0.25 (SD=0.25). These

parameters are based on the difficulty levels from an operational MST item bank used by the Medical Council of Canada. We simulated more variation among the items in the easy level because, in our experience, easy items are more readily created and hence more common in an item bank. Moreover, the operational MST item bank from the Medical Council of Canada contained more variation among the easier items relative to medium and hard items.

Examinees were routed in the simulated MST using a simple strategy determined by their number correct score. Examinees who answered all four items correctly moved to a harder module; examinees who scored 0 or 1 moved to an easier module; examinees who scored 2 or 3 moved to a module with the same difficulty level. The bank size for our MST simulation was fixed, as it contained 100 items at each difficulty level, with 90 non-DIF and 10 DIF items per level. The items were randomly selected at each difficulty level during the adaptive test administration. Examinees' ability estimates in the MST simulation were generated from a bimodal distribution to ensure equal participation (i.e., exposure) rates across the three difficulty levels. In other words, the same distribution was used to generate examinees in the reference and focal groups meaning that there was no systematic difference in ability between these two groups (this condition is often referred to *no impact* in the DIF literature; see, for example, Lei et al., 2006, p. 248). By combining the examinee generation process using the bimodal distribution with the number-correct routing strategy, an expected use rate could be anticipated for each item. For the condition with 3000 examinees (1500 examinees in both the reference and focal groups), the adaptive process combined with our fixed-length item bank yields approximately 100-175 examinees per group who responded to each item. That is, 3000 examinees were administered 28 items adaptively using a bank of 300 items. With the bank containing 100 items per difficulty level, this resulted in an exposure rate ranging from 6.67% to 11.67% per item. With examinees from the reference and focal groups each having an equal chance of receiving an item from the bank, this resulted in 100-175 examinees per group who responded to each

item (i.e., 6.67% x 3000 / 2 = 100; 11.67% x 3000 / 2 = 175). For the condition with 6000 examinees (3000 examinees in both the reference and focal groups), the adaptive process combined with our fixed-length item bank yields approximately 200-300 examinees per group who responded to each item. For the condition with 9000 examinees (4500 examinees in both the reference and focal groups), the adaptive process combined with our fixed-length item bank yields approximately 300-450 examinees per group who responded to each item.

Sample size, then, serves as the second independent variables in our study—it may also be one of the most important variables to consider when computing DIF in CAT because sample sizes are typically very small, especially compared to paper-based tests where every item is administered to all examinees. In CAT, the number of examinees who write any one item is limited because item exposure rates are often kept to a minimum to ensure test security. As a result, DIF methods designed to help monitor bias must function in testing environments where the total number of items in the bank may be large but the number of examinees who respond to any one of those items is relatively small. For instance, if 1000 examinees write an MST, but the item exposure rate is 10%, then each item would only be administered to 100 examinees. The problem is compounded when these 100 examinees are further divided into the reference and focal groups, and then separated by ability prior to computing the DIF statistic. Three different sample sizes were evaluated in the current study—small (i.e., 100-175 examinees per group per item), moderate (i.e., 200-300), and large (i.e., 300-450). We also evaluated the consequences of balanced (i.e., same sample size range in the reference and focal groups) and unbalanced (i.e., different sample size range in the reference and focal groups) sample sizes on DIF detection, as this was an important variable that affected CATSIB performance in the Lei et al. (2006) study.

Dependent Variables

Two types of dependent variables were used to evaluate the performance of CATSIB: Type I error and power. Type I error refers to the probability that CATSIB will incorrectly identifying an item as displaying DIF when, in fact, it does not. We can call this concept a "false alarm", meaning a non-DIF item is falsely signalled as a DIF item. Type I error was calculated across 100 simulated analyses for the 270 non-DIF items (i.e., of the 300 items in the bank, only 10 DIF items were simulated at each of the three difficulty levels). Ideally, when an item is specified as a non-DIF item, the probability of detecting this item mistakenly as a DIF item (i.e., Type 1 error) should be close to 0%. In practice, however, we never use a test that maintains a Type I error rate of 0% because this outcome would adversely affect power, given that Type I error and power depend on one another (see Figure 3). Hence, we use the nominal level of 0.05 or 5% to assess the Type I error performance of the non-DIF items in our study. This 5% level corresponds to a $\hat{\beta}_{UNI}$ value of 0.08.

Power refers to the probability of correctly detecting a DIF item. We call this concept a "correct decision" because we are identifying DIF items correctly. Power was calculated across 100 simulated analyses for the 30 DIF items. Ideally, the probability of detecting a DIF item correctly should be 100%. Unfortunately, high power rates can also lead to an unreasonably high Type I error rates (i.e., all items are identified as DIF items regardless of their true designation as DIF or non-DIF). For this study, a $\hat{\beta}_{UNI}$ value of 0.08 is used as the criterion for identifying a DIF item, meaning that items with $\hat{\beta}_{UNI}$ = 0.08 or greater are considered DIF items. That is, by generating a normal distribution of DIF items where the $\hat{\beta}_{UNI}$ has a mean of 0.11 and a standard deviation of 0.04, the criterion of 0.08 is used to identify DIF items because this value is two standard deviations above the null distribution of no DIF (see Figure 3) and would represent a large difference between the performance of the reference and focal group

examinees. A nominal level of 0.80 or 80% is used to assess the power of CATSIB to correctly detect DIF items in our study. This 80% level corresponds to a $\hat{\beta}_{UNI}$ value of 0.11.

RESULTS

We present the results in two sections. First, we present the results for the balanced sample size conditions. The Type I error and power rates as a function of the item difficult level in each module are described. Second, we present the results for the unbalanced sample size conditions.

Balanced Sample Size

With a small (100-175) number of examinees per group, the Type I error differed across the three difficulty levels (see Table 2). The rate for the easy items was 0.17; the rate for the medium items was 0.22; the rate for the hard items was 0.21. Across all three difficulty levels, the overall Type I error rate was 0.20. The power rates also differed across the three difficulty levels. Power for the easy items was 0.74; power for the medium items was 0.72; power for the hard items was 0.74. The overall power rate was 0.73. These results reveal that when sample sizes are small, both the Type I error and power rates fail to meet our acceptable level of performance of 0.05 and 0.80, respectively.

With a moderate (200-300) number of examinees per group, the Type I error and power rates also differed across the three difficulty levels and improved, relative to the results in the small sample condition. The Type I error for the easy items was 0.05; the rate for the medium items was 0.08; the rate for the hard items was 0.07. The overall Type I error rate was 0.07. Power for the easy items was 0.80; power for the medium items was 0.80; power for the hard items was 0.78. The overall power rate was 0.79. These results, although better than the previous condition, again failed to meet our acceptable level of performance of 0.05, and 0.80, respectively, except for the easy items. Power did, however, either meet or approach our criterion of 0.80 for the medium and overall difficulty levels.

With a large (300-475) number of examinees per group, the Type I error and power rates produced the most acceptable results in our simulation study and met both our criteria. The Type I error for the

easy items was 0.02; the rate for the medium items was 0.03; the rate for the hard items was 0.03. The overall Type I error rate was 0.02. Power for the easy items was 0.85; power for the medium items was 0.84; power for the hard items was 0.84. The overall power rate was 0.84.

Unbalanced Sample Size

With a small focal sample (100-175 examinees per group) and a moderate reference sample (200-300 examinees per group), the Type I error rates, again, differed across the three difficulty levels (see Table 3). The rate for the easy items was 0.10; the rate for the medium items was 0.15; the rate for the hard items was 0.14. The overall Type I error rate was 0.13. The power rates also differed across the three difficulty levels. Power for the easy items was 0.76; power for the medium items was 0.82; power for the hard items was 0.80. The overall power rate was 0.79. These results reveal that with a small focal and moderate reference sample, the Type I error rates fail to meet the acceptable criterion of 0.05, but the power rate was acceptable for the medium and hard difficulty levels and approached the criterion of 80% for the overall rate.

With a small focal sample (100-175 examinees per group) and a large reference sample (300-475 examinees per group), the Type I error rates remained below the acceptable standard of performance whereas the power rate approached the criterion, but only for the easy difficulty level. The Type I error for the easy items was 0.08; the rate for the medium items was 0.13; the rate for the hard items was 0.13. The overall Type I error rate was 0.11. Power for the easy items was 0.79; power for the medium items was 0.75; power for the hard items was 0.77. The overall power rate was 0.77.

With a moderate focal sample (200-300 examinees per group) and a large reference sample (300-475 examinees per group), the Type I error rates and power rates met the acceptable criterion of 0.05 and 0.80 for all difficulty levels. The Type I error for the easy items was 0.03; the rate for the medium items was 0.05; the rate for the hard items was 0.05. The overall Type I error rate was 0.04. Power for

the easy items was 0.85; power for the medium items was 0.80; power for the hard items was 0.82. The overall power rate was 0.82.

A summary of the overall results (i.e., results across the three difficulty levels), relative to our acceptable Type I error and power criteria, is presented in Table 4. The Type I error criterion of 0.05 was met for two of the six sample size conditions—large reference/moderate focal and large reference/large focal. The power criterion of 0.80 was also met in two of the MST sample size conditions—large reference/moderate focal and large reference/large focal. The results from the moderate reference/moderate focal approached the criterion, with an overall power result of 0.79.

CONCLUSIONS AND DISCUSSION

The purpose of this study was to evaluate the performance of CATSIB for detecting DIF when items in both the matching and studied subtest were administered adaptively in the context of a realistic multi-stage adaptive test. Our research contributes to the study of DIF in CAT in two ways. First, we evaluated CATSIB when items on the studied and matching subtests were administered adaptively. Nandakumar and Roussos (2001, 2004) and Lei et al. (2006) only evaluated CATSIB when items on the matching subtest were administered adaptively. Second, we evaluated CATSIB in the realistic and, increasingly, popular CAT context of multi-stage adaptive testing (MST). To-date, there has been no research on DIF in MST. We simulated MST using a four-item module in a seven-panel administration, where the first panel contained a single four-item module that was common to all examinees. The second to seventh panels each contained three modules with four items per module at three different difficulty levels. Each examinee wrote seven modules thereby completing 28 items.

Thee independent variables were manipulated: item difficulty, sample size, and balanced/unbalanced design. Because item difficulty is a prominent variable in MST that could affect CATSIB DIF detection, it served as the first independent variable in this study. Module difficulty level was characterized as easy, medium, and hard, and was based on the empirical results from an

operational MST program. Sample size served as the second independent variable in our study because the number of examinees who write any one item may be quite small in CAT. Three different sample sizes were evaluated—small (i.e., 100-175 examinees per group who responded to each item), moderate (i.e., 200-300), and large (i.e., 300-450). The design—balanced versus unbalanced samples in the reference and focal group—was shown by Lei et al. to affect CATSIB DIF detection. Therefore, the consequences of balanced (i.e., same sample size range in the reference and focal groups) and unbalanced (i.e., different sample size range) sample sizes on CATSIB DIF detection was also evaluated.

Two types of dependent variables were used to evaluate the performance of CATSIB, Type I error and power. Type I error refers to the probability that CATSIB will incorrectly identifying an item as displaying DIF when, in fact, it does not. Type I error was calculated across 100 simulated analyses for the 270 non-DIF items. We used the nominal level of 5% to assess the Type I error performance of the non-DIF items. Power refers to the probability of correctly detecting a DIF item. Power was calculated across 100 simulated analyses for the 30 DIF items. A nominal level of 80% was used to assess the power of CATSIB to correctly detect DIF items.

CATSIB met the acceptable criteria, meaning that the Type I error and power rates met 5% and 80%, respectively, for the large reference/moderate focal sample and the large reference/large focal sample conditions. The criteria were also met in only one other study condition—easy difficulty level with a moderate number of reference and focal group examinees. These results indicated that CATSIB can be used to consistently and accurately detect DIF on an MST, but only with moderate to large samples. In other words, CATSIB will identify DIF items with adequate Type I error protection and power across a range of module difficulty levels when a minimum sample size of 475 examinees (i.e., 175 focal group + 300 reference group) is used. CATSIB performs even better, meaning lower Type I errors and higher power, with 600 examinees (i.e., 300 examinees in the focal and reference groups).

Directions for Future Research

Although there is little research on the effects of small sample size DIF detection with CATSIB, the studies by Nandakumar and Roussos (2001, 2004) and Lei et al. (2006) provide an important point of reference for the current study. Recall that Nandakumar and Roussos and Lei et al. both used a pretest design where single items in the matching subtest were administered adaptively but items on the studied subtest were not. The current study, by comparison, assessed modules of items in the matching and studied subtest where the items in both subtests were administered adaptively. To evaluate Type I error, Nandakumar and Roussos used six non-DIF items with b-parameters ranging from -2 to 2, a-parameters ranging from 0.5 to 1.7, and c-parameters ranging from 0.12 to 0.22. Using a two-tailed test with no ability differences between groups (i.e., no impact), the average Type I error across the six items was 0.05 with 250 reference/250 focal group examinees, 0.05 with 500 reference/250 focal group examinees, and 0.06 with 500 reference/500 focal group examinees. Lei et al. used 16 non-DIF items with b-parameters ranging from -1.95 to 1.95, a-parameters ranging from 0.74 to 1.5, and c-parameters fixed at 0.15. Using a two-tailed test with no ability differences between groups, the average Type I error across the 16 items was 0.06 with 500 reference/500 focal group examinees. In the current study, b-parameters in four-item modules at three difficulty levels ranged from -1.25 to 0.25, a-parameters were generated with a mean of 0.80 and a standard deviation of 0.20, and c-parameters were fixed at 0. Using a two-tailed test with no ability differences between groups, the average Type I error across the module difficulty levels was 0.07 with 200-300 reference/200-300 focal group examinees, 0.04 with 300-475 reference/200-300 focal group examinees, and 0.02 with 300-475 reference/300-475 focal group examinees. In short, the CATSIB Type I error rates are quite consistent across the three studies (i.e., Type I error rates ranged from 0.02 to 0.07).

The outcomes across the three studies are also comparable, generally speaking, for the power rates. To evaluate power, Nandakumar and Roussos assessed medium and large DIF items. We focus only on the results for the six large (i.e., $\hat{\beta}_{UNI}$ = 0.10) DIF items so comparisons can be made with the current study. The average power across these six DIF items was 0.73 with 250 reference/250 focal group examinees, 0.84 with 500 reference/250 focal group examinees, and 0.94 with 500 reference/500 focal groups examinees. Lei et al. also evaluated large DIF in one set of their study conditions. Using eight large DIF items, meaning the area between the unidirectional item characteristic curves for the reference and focal groups was set to 0.60, the average power was 1.00 with 500 reference/500 focal group examinees. In the current study, the average power for the 30 large DIF items across 100 replication in the modules was 0.79 with 200-300 reference/200-300 focal group examinees, 0.82 with 300-475 reference/200-300 focal group examinees, and 0.84 with 300-475 reference/300-475 focal group examinees. In sum, CATSIB can effectively identify large DIF items across a variety of sample size conditions. Power rates, across the three studies, ranged from a low of 0.73 for samples with 500 examinees in total to 1.00 for samples with 1000 examinees in total. The consistent Type I error and power results between Nandakumar and Roussos and Lei et al. with the current study also provide corroborating evidence to support our recommendation that samples as small as 475 examinees in total (i.e., 175 focal group + 300 reference group) can be used to identify large DIF items in CAT with CATSIB, but that 600 examinees in total (i.e., 300 examinees in the focal and reference groups) is preferred. More research, however, should be conducted using modules with more varied and diverse difficulty levels to evaluate CATSIB in other MST conditions commonly found in operational testing programs, given that module difficulty influences DIF detection. Our results, although based on the outcomes from an operational testing program, were restricted to the module difficulty range of -1.25 to 0.25. We also focused on conditions where the ability distribution between the reference and focal groups were comparable (i.e., no impact). Future studies should also evaluate

CATSIB DIF detection in the presence of impact, meaning when the ability distribution between the two groups are markedly different.

A second area for future research resides directly with the CATSIB statistic itself. Research should be conducted to evaluate the potential to improve the accuracy of the examinees' regression-corrected ability estimate (θ^*) by using the standard error of measurement (SEM) conditional on each examinee rather than the overall SEM. In the current study, the θ^*s were estimated based on the regression correction formula for Shealy and Stout's (1993) SIBTEST, in which the overall SEM is assumed to be identical across the examinees in focal and reference groups. The overall SEM was also used by Nandakumar and Roussos (2001, 2004) in their initial description and evaluation of CATSIB. The use of an overall SEM, however, may lead to a potential limitation—examinees in the same group (either focal or reference) are assumed to share the same error variance, even when they possess different ability levels. For instance, when an examinee with a high ability level answers a difficult item, it tends to yield a small SEM for that examinee. Conversely, when an examinee with low ability level attempts the same item, it yields a large SEM for that examinee. In this example, the ability estimates (θ^*s) may be biased because the current version of CATSIB applies the overall SEM to both examinees. To correct for this bias, Raju, Price, Oshima, and Nering (2007) proposed a procedure which can estimate the SEM for each examinee, conditional on their ability level. Additional research should be conducted to combine the CATSIB procedure described in this study with Raju et al.'s (2007) procedure to evaluate the performance of a *modified CATSIB* approach. A modified CATSIB procedure is expected to yield more accurate SEMs and θ^*s for each examinee, thereby leading to a decrease in Type I error and an increase in power which means improved statistical DIF detection in CAT.

A third area of research should focus on using CATSIB with the Roussos and Stout (1996a) DIF analysis paradigm to understand why DIF occurs in CAT. DIF studies are undertaken for many reasons. One reason to conduct a DIF study may be to identify and remove items that elicit large group

differences, even when the reasons for these differences are not apparent. Another reason to conduct a DIF study is to better understand the nature of these group differences. The DIF analysis paradigm is used to address the second reason. Roussos and Stout (1996a) developed this framework to unify the substantive and statistical analyses because many researchers and practitioners reported that the outcomes from DIF statistical analyses alone were not interpretable. It serves as one of the very first model-based approaches for identifying and interpreting the factors that elicit group differences. Because the independent variables that affect CATSIB's statistical performance are becoming more apparent, researchers and practitioners can also begin to use this psychometric procedure to start to evaluate the factors that could explain the presence of DIF. Zwick (2000, 2010) claimed, for instance, that computer-based test administration might elicit several new and important sources of DIF that are not present in paper-based testing, including differential computer familiarity, facility, and anxiety. These hypotheses can now be assessed, at least in some CAT conditions, by using CATSIB with the DIF analysis paradigm to study the factors that could explain why DIF occurs.

REFERENCES

Applied Measurement in Education (2006). *Special issue: An introduction to multistage testing*, 19, 185-260.

Bartram, D. (2006). Testing on the internet: Issues, challenges, and opportunities in the field of occupational assessment. In D. Bartram & R. Hambleton (Eds.), *Computer-based testing and the internet* (pp. 13-37). Hoboken, NJ: Wiley.

Breithaupt, K. J., Mills, C. N., & Melican, G. J. (2006). Facing the opportunities of the future. In D. Bartram & R. Hambleton (Eds.), *Computer-based testing and the internet* (pp. 219-251). Hoboken, NJ: Wiley.

Bennett, R. (2001). How the internet will help large-scale assessment reinvent itself. *Educational Policy Analysis Archives, 9,* 1-23.

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items.* Newbury Park: Sage.

Drasgow, F., & Mattern, K. (2006). New tests and new items: Opportunities and issues. In D. Bartram & R. Hambleton (Eds.), *Computer-based testing and the internet* (pp. 59-76). Hoboken, NJ: Wiley.

Educational Testing Service (2007). *Innovations: News on Research, Products, and Solutions for Learning and Education,* Summer 2007. Princeton, NJ: Educational Testing Service.

Educational Testing Service (2009). *ETS Guidelines for Fairness Review of Assessments.* Princeton, NJ: Educational Testing Service.

Englehard, G,. Hansche, L., & Rutledge, K. E. (1990). Accuracy of bias review judges in identifying differential item functioning on teacher certification tests. *Applied Measurement in Education, 3,* 347-360.

Gierl, M. J. (2005). Using a dimensionality-based DIF analysis paradigm to identify and interpret constructs that elicit group differences. *Educational Measurement: Issues and Practice, 24,* 3-14.

Gierl, M. J., Bisanz, J., Bisanz, G., Boughton, K., & Khaliq, S. (2001). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests. *Educational Measurement: Issues and Practice, 20,* 26-36.

Gierl, M. J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests. *Journal of Educational Measurement, 38,* 164-187.

Gierl, M. J., Bisanz, J., Bisanz, G., & Boughton, K. (2003). Identifying content and cognitive skills that produce gender differences in mathematics: A demonstration of the DIF analysis framework. *Journal of Educational Measurement, 40,* 281-306.

Lei, P. W., Chen, S. Y., & Yu, L. (2006). Comparing methods of assessing differential item functioning in a computerized adaptive testing environment. *Journal of Educational Measurement, 43,* 245-264.

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

Nandakumar, R., & Roussos, L. (2001, July). *CATSIB: A modified SIBTEST procedure to detect differential item functioning in computerized adaptive tests.* Law School Admission Council Computerized Testing Report 97-11.

Nandakumar, R., & Roussos, L. (2004). Evaluation of the CATSIB DIF procedure in a pretest setting. *Journal of Educational and Behavioral Statistics, 29*, 177-199.

O'Neill, K. A., & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 255-276). Hillsdale, NJ: Lawrence Erlbaum.

Plake, B. S. (1980). A comparison of a statistical and subjective procedure to ascertain item validity: One step in the validation process. *Educational and Psychological Measurement, 40,* 397-404.

R Development Core Team (2011). *An Introduction to R.* Version 2.13.0.

Raju, N. S., Price, L. R., Oshima, T. C., & Nering, M. L. (2007). Standardized conditional *SEM*: A case for conditional reliability. *Applied Psychological Measurement*, *31*, 169 – 180.

Roussos, L., & Stout, W. (1996a). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20,* 355-371.

Roussos, L., & Stout, W. (1996b). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement, 33,* 215-230.

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 58,* 159-194.

Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp.329-348). Mahwah, NJ: Erlbaum.

Standards for Educational and Psychological Testing. (1999). Washington, DC: American Educational Research Association, American Psychological Association, National Council on Measurement in Education.

van der Linden, W. J., & Glas, C. A. W. (2000). *Computerized Adaptive Testing: Theory and Practice*. Dordrecht, The Netherlands: Kluwer.

van der Linden, W., & Glas, C. A. W. (2010). *Elements of adaptive testing*. New York: Springer.

Zenisky, A., Hambleton, R., & Luecht, R. (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.

Zenisky, A. L., & Sireci, S. G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education, 15,* 337-362.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.) *Differential item functioning* (pp. 337-347). Hillsdale, NJ: Erlbaum.

Zwick, R. (2000). The assessment of differential item functioning in computer adaptive tests. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized Adaptive Testing: Theory and Practice* (pp. 221-244). Dordrecht, The Netherlands: Kluwer.

Zwick, R. (2010). The investigation of differential item functioning in adaptive tests. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of Adaptive Testing* (pp. 331-352). New York: Springer.

Table 1.

The Generated Item Difficulty Parameters for the Easy, Medium, and Hard MST Modules

	a-parameter		b-parameter		Number of Items	
Difficulty Level	Mean	SD	Mean	SD	Non-DIF	DIF
Easy	0.80	0.20	-1.25	0.50	90	10
Medium	0.80	0.20	-0.25	0.25	90	10
Hard	0.80	0.20	0.25	0.25	90	10

Table 2.

The Type I Error and Power Rates for the Balanced Sample Size Conditions

Focal/Reference Group Size	Difficulty Level	Type 1 Error	Power
Small (100-175)/Small (100-175)	Easy	0.17	0.74
	Medium	0.22	0.72
	Hard	0.21	0.74
	Total	0.20	0.73
Moderate (200-300)/Moderate (200-300)	Easy	0.05	0.80
	Medium	0.08	0.80
	Hard	0.07	0.078
	Total	0.07	0.79
Large (300-475)/Large (300-475)	Easy	0.02	0.85
	Medium	0.03	0.84
	Hard	0.03	0.84
	Total	0.02	0.84

Table 3.

The Type I Error and Power Rates for the Unbalanced Sample Size Conditions

Focal/Reference Group Size	Difficulty Level	Type 1 Error	Power
Small (100-175)/Moderate (200-300)	Easy	0.11	0.76
	Medium	0.15	0.82
	Hard	0.14	0.80
	Total	0.13	0.79
Small (100-175)/Large (300-475)	Easy	0.09	0.79
	Medium	0.13	0.75
	Hard	0.13	0.77
	Total	0.11	0.77
Moderate (200-300)/Large (300-475)	Easy	0.03	0.85
	Medium	0.05	0.80
	Hard	0.05	0.82
	Total	0.04	0.82

Table 4.

Summary of Sample Size Results Relative to the Type I Error and Power Criteria of 5% and 80%, Respectively

	Focal Group					
	Small (100-175)		Moderate (200-300)		Large (300-475)	
Reference Group	Type 1 Error	Power	Type I Error	Power	Type 1 Error	Power
Small (100-175)	NO	NO	--	--	--	--
Moderate (200-300)	NO	NO	NO	NO	--	--
Large (300-475)	NO	NO	YES	YES	YES	YES

Figure 1. An example of a three-panel, seven-module multistage adaptive test.

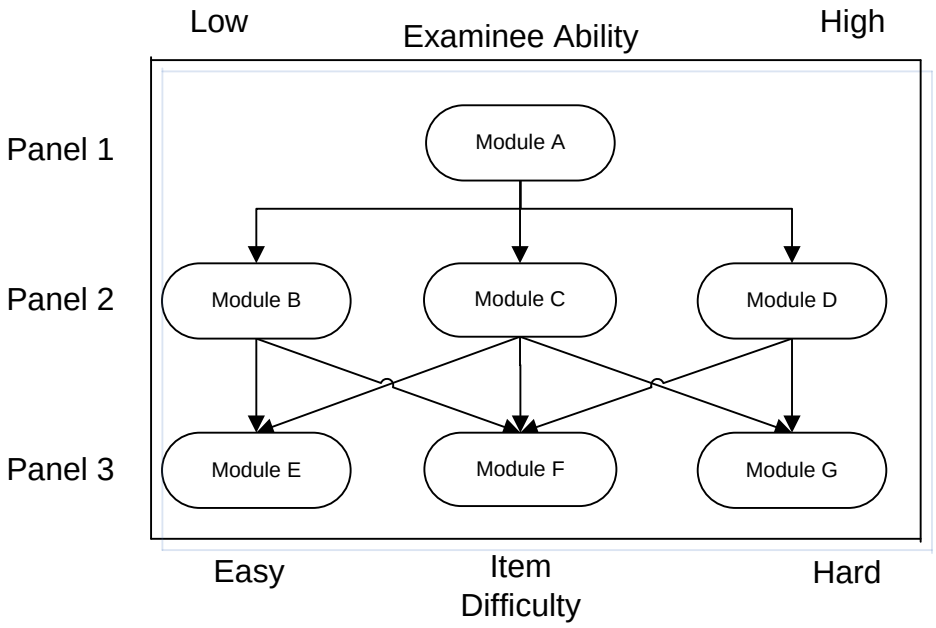

Figure 2. The simulated MST used in the current study: A nineteen-module, seven-panel MST with four-items per module.

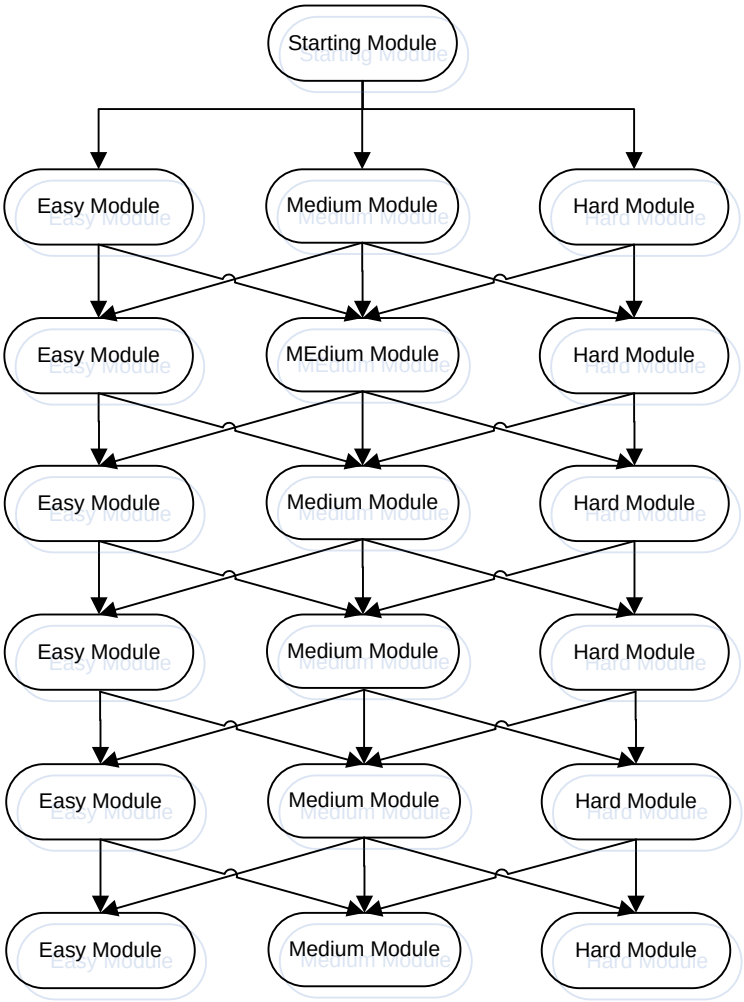

Figure 3. A summary of the null and alternative distributions required to set the nominal levels of Type I error and power in our study. The region in the null distribution represents the nominal level of Type I error (i.e., 5%) whereas the region in the alternative distribution indicates the nominal level of power (i.e., 80%).