Detecting Biased Items Using CATSIB to Increase Fairness in Computer Adaptive Tests

Man-Wai Chu, Hollis Lai

Centre for Research in Applied Measurement and Evaluation, University of Alberta

In educational assessment, there is an increasing demand for tailoring assessments to individual examinees through computer adaptive tests (CAT). As such, it is particularly important to investigate the fairness of these adaptive testing processes, which require the investigation of differential item function (DIF) to yield information about item bias. The performance of simultaneous item bias test for computer adaptive test (CATSIB), a revision of simultaneous item bias test (SIBTEST) to accommodate CAT responses, in detecting DIF in a multi-stage adaptive testing (MST) environment is investigated in the present study. Specifically, the power and type I error rates on directional DIF detection of an MST environment when positive and negative impact, group performance differences, were investigated using simulation procedures. The results revealed that CATSIB performed relatively well in identifying the items with DIF when characteristics of the group and items were known. Assessment stakeholders are able to use these results to enhance test items, which provide students with fair and equitable adaptive testing environments.

Dans le domaine de l'évaluation pédagogique, il existe une demande croissante pour personnaliser les évaluations par l'adoption d'examens informatisés adaptatifs (CAT : computer adaptive test). Ainsi, il est important de se pencher sur l'équité de ces processus adaptatifs appliqués aux évaluations; ceci exige la détection du fonctionnement différentiel d'items de sorte à déterminer le biais d'item. L'étude présente porte sur le rendement de la procédure CATSIB, et une révision de la procédure SIBTEST pour accommoder les réponses aux examens informatisés adaptatifs et détecter le fonctionnement différentiel d'items dans un environnement d'examens adaptatifs à plusieurs étapes. Les résultats indiquent que la procédure CATSIB a fonctionné relativement bien dans l'identification des items avec un fonctionnement différentiel quand les caractéristiques du groupe et des items étaient connues. Les intervenants en évaluations pourront se servir de ces résultats pour améliorer les items des évaluations de sorte à offrir aux étudiants des examens adaptatifs équitables.

"That test was not fair!" Many students have uttered this simple phrase at some point during their time in school. Students often question the notion of fairness after receiving disappointing test results that may have dire consequences. Fairness explores many considerations in relation to the goals of the test, one of which is achieving equality of opportunities among students (AERA, APA, & NCME, 1999). Test results are frequently used to make key decisions on students' futures, such as selection into remedial or advanced classes, admission to postsecondary institutions, or attaining scholarships. Tests are expected to be representative of a program-of-study in assessing how much knowledge, skills, and attributes a student has acquired (Cizek, 2009; Mislevy, Steinberg, & Almond, 2002). Thus, each test has a set of items that represents a portion of the program-of-study and accounts for a fraction of each student's mark. Since each test is comprised of a finite number of items, it is important for each item to be fair in assessing students equitably (AERA, APA, & NCME, 1999). Therefore tests, especially high stakes tests, require items that are all fair and equitable, meaning they measure students' knowledge free of irrelevant factors such as gender or cultural bias (Camilli, 2006). The general purpose of this paper is to investigate the detection of unfair items in a test using a computer program, Simultaneous Item Bias test for Computer Adaptive Test (CATSIB). The findings from this study will be useful to testing agencies because it will inform their practices of detecting fair and equitable items in all tests administered to students.

Differential Item Functioning

The idea of fair and equitable tests is well documented in the realm of education as researchers have devoted decades of studies towards perfecting test development (AERA, APA, & NCME, 1999; Brennan, 2006; Downing & Haladyna, 2006). Tests that are fair and equitable allow educators to assess students with a variety of backgrounds. Issues of test fairness and equality tend to deal with test items being biased against particular groups of students (Camilli, 2006; Gierl, Rogers, & Klinger, 1999). Specifically, items that elicit a systemic preference towards a specific group that is not related to their performance, or seen as unfair towards particular groups of students, are called biased (Hambleton, Swaminathan, & Rogers, 1991). Moreover, item bias is a serious concern for test developers and users because they are often difficult to remedy. As a result, biased items yield test results with systematic errors that distort the inferences made for members of a particular group, such as females, Aboriginals, or French-speaking examinees (Zumbo, 1999). Consider the following example of a biased test item from the Alberta grade 6 Mathematics Achievement Test in Gierl et al.'s paper (1999), which was written in English and translated into French for a provincial exam:

> On the first day of filming, the crew arrived on the set at 5:20 A.M. They left the set at 8:15 P.M. How long did the crew spend on the set that day?
>
> A. 3 h 5 min
> B. 5 h 5 min
> C. 13 h 35 min
> D. 14 h 55 min
>
> Le premier jour du tournage, l'équipe arrive au plateau de projection à 5 h 20 du matin. Elle quitte le plateau à 20 h 15. Combien de temps l'équipe est-ce que l'équipe passe sur le plateau le premier jour?
>
> A. 3 h 5 min
> B. 5 h 5 min
> C. 13 h 35 min
> D. 14 h 55 min

The two forms of this item were deemed non-equivalent, as the English form contained a 12-hour clock with AM and PM whereas the French form used a 24-hour clock, which is consistent with French culture. Since students are required to interpret the difference between the AM and PM time, the French form's use of the 24-hour clock made the correct response more apparent.

To identify these biased items on large-scale assessments, groups of content specialists or sensitivity committees review the test items to ensure every item is fair and equitable prior to administration of the assessment. However, the subjective nature of detecting biased items using content specialists often produce unreliable results because of individual differences (Camilli & Shepard, 1994). For example, a specialist may identify biased items based on two genders, but it is more difficult when considering multiple ethnicities because each group needs to be considered to ensure the item is fair for everyone.

Additionally, a practical issue for not relying solely on content specialists to review the whole test for biased items is the high cost of hiring them. As such, it is not financially feasible to have the specialists review every item of the test. Instead, it is important to have the specialists review only a few of the biased items that have already been identified as problematic through the use of a computer program. For example, after the assessment is administered differential item functioning (DIF) analyses can be conducted to identify whether there are any biased or DIF items. The computer program identifies the DIF items when there are differences in response patterns between two groups of students (Zumbo, 1999). The content specialists then review the DIF items to confirm whether or not they are in fact problematic. Thus, a statistical approach of detecting item bias using a computer program can lower the costs of item development since specialists are directed to review a few specific items instead of the whole test (Roussos & Stout, 1996). Additionally, this statistical approach also provides an additional line of evidence to indicate an item is biased. Although these statistical programs are economical for testing companies, they have limitations such as their inability to identify the reasons behind why specific items are considered to be DIF. These programs are able to detect DIF items in many assessment formats including computer-based tests.

Computer-Based Tests

As advancements in computer technology have proven beneficial for large-scale assessments (Bennett, 2006), many paper-and-pencil format large-scale assessments are now administered electronically through the Internet (Lightstone & Smith, 2009). The benefits of administering electronic assessments is the ability to include innovative item types that integrate digital media to increase the type of knowledge and skills to be measured (Bartram, 2006; Breithaupt, Mills, & Melican, 2006; Scalise & Gifford, 2006; Sireci & Zenisky, 2006; Zenisky & Sireci, 2002) and to personalize assessments to individual examinees through the use of Computer Adaptive Tests (CAT) (Drasgow, Luecht, & Bennett, 2006; Weiss, 1982). Thus, instead of administering the same set of items to all examinees, CAT exams can be personalized such that all students are capable of completing the exam and provide better performance estimates. That is, personalization of the tests allow more items that target students' ability levels to be administered so that more evidence may be collected to justify their level of achievement.

The use of CAT exams in education has been studied and debated for many years. For example, supporters of CAT argued that adaptive tests lead to more accurate measurement of students; however, their opponents disputed that these tests are not equal in the strictest statistical sense because the same items are not administered to all students (Way, Twing, Camara, Sweeney, Lazer, & Mazzeo, 2010). While the philosophical debate regarding whether or not to use CAT exams in education is ongoing, the focus of this paper is to propose a method of detecting unfair items within CAT.

The adaptability of CAT streamlines examinees toward certain items depending on their actual knowledge and skills, known as their performance levels. Examinees are presented with items tailored for their performance levels, which are calculated based on the responses given to previous items. For example, if examinees are given an item of moderate difficulty and they get the item correct, they are then presented with an item of higher difficulty. This continues until the examinee's performance has been determined based on the stopping rule of CAT (Gershon, 2005). Adapting the difficulty of each item to the examinee's performance allows the test to better assess the examinee's performance, while shortening the length of the test when compared to paper-and-pencil tests that yield the same precision (Weiss, 1982). In other words, adaptive tests will provide a more precise estimate on examinee abilities across a different range of performance. For example, in a paper-based test, if simple items are administered to high performing students they will most likely get those items correct, but those results will simply indicate the student is able to answer simple items and does not indicate how much the student actually knows. In contrast, if difficult items were administered to high performing students, they would provide a more accurate estimate of their abilities because each item would find the cut between what the students know and do not know, determining their performance level. One form of CAT, which has gained popularity, is multi-stage testing (MST).

Multi-Stage Testing

MST is a unique form of CAT where modules of items, also known as testlets, are administered to examinees based on their performance levels. The modules administered are determined based on the responses given in previous modules. For example, if the examinee performs well on the first module, then the computer will provide the examinee with a more difficult module of items. The difficulty of a module is designed to match examinees' performance levels in order to optimally estimate the examinees' performance with minimal tolerance of error (Mean, 2006). MST is used in several large-scale examinations such as the American Institute of Certified Public Accountants (CPA) Uniform Examination and the Medical Council of Canada Qualifying Examination (American Institute of CPAs, 2012; Gierl, Lai, & Li, 2011; Medical Council of Canada, 2012).

The key benefit of administrating exams in a MST environment is the idea of grouping items into modules, which can potentially allow the examiner to administer items adapted to the examinee's performance while meeting program-of-study content posed constraints (see Gierl, Lai, & Li, 2013). Items from the same unit can be grouped into different modules and administered at each stage of the exam, ensuring a breadth of content is tested. Moreover, since MST only requires the performance level to be estimated at the end of each module, the processing power required would be a fraction of that necessary for CAT resulting in fewer hardware requirements and financial resources to implement (Chang & Ying, 1999).

In a MST environment, the number of examinees writing any particular item significantly decreases because examinees are streamed towards different modules of items. The low sample size of both the reference and focal groups creates challenges in detecting items with DIF because each group may typically provide 100-300 examinees for each item. Typically traditional programs used to detect DIF items, such as Mantel-Haenszel (MH) or Simultaneous Item Bias Test (SIBTEST), do not perform well when there is less than 200 examinees that respond to an item (Gotzmann & Boughton, 2004; Roussos & Stout, 1996). As mentioned earlier different examinees write a different set of items, which complicates the comparison of examinees because there may be no common items to compare the examinees with.

Nandakumar and Roussos (2004) adjusted SIBTEST to produce a new program, CATSIB, which is able to detect DIF when examinees write different sets of items.

Simultaneous Item Bias Test for Computer Adaptive Tests

CATSIB and SIBTEST utilize similar statistical calculations (Nandakumar & Roussos, 2004; Gierl et al., 2011; Li, Gierl, & Lai, 2013; Shealy & Stout, 1993; Van der Linden & Glass, 2010). The main difference between CATSIB and SIBTEST is an IRT based regression correction equation. CATSIB calculates the estimated performance levels of the examinees through a regression-corrected Item Response Theory (IRT)-based performance estimate to adjust the means of the reference and focal subgroups (Nandakumar & Roussos, 2004). IRT is a method for designing, analysing, and scoring a test that measures students' performance. This method rests on two basic postulates: (a) a set of factors can explain a students' performance on tests (i.e., traits) and (b) relationship between students' performance on each item and the set of factors can be described using a mathematical function (please refer to Hambleton et al., 1991). SIBTEST uses a regression correction equation based on the standard error of examinees' performance levels (Li et al., 2013). Reference and focal groups are the names of the two groups of examinees being studied. For example, examinees writing the English form of a test may be called the reference group while the examinees writing the French form would be called the focal group. Once the two groups have been matched on performance, the probability of each group correctly answering an item is calculated and subtracted (DIF(θ) = $P_R(\theta)$ – $P_F(\theta)$; Gierl et al., 2011; Roussos & Stout, 1996; Shealy & Stout, 1993). The DIF(θ) values are then summed and weighted overall performance levels, signified by θ. Next, a comparison of these DIF(θ) values, the difference between the correct responses or weighted mean differences, is done to determine the presence of DIF. Using this statistical procedure, it is possible for examinees at the extreme ends of the performance distribution, very high and very low performance examinees, to regress more than examinees with average performance levels (Li et al., 2013). Thus, it is important to investigate how different performance levels, known as impact, affect CATSIB's performance in detecting DIF items.

The Effects of Impact on DIF Detection

Varying performance levels that reflect actual knowledge and experience differences between two intact groups is known as impact (Dorans & Holland, 1993). There are often systematic differences in performance, or impact, between two groups of examinees. This can be seen on a typical SAT-Mathematics item where the performance of International Baccalaureate (IB) students tends to be higher than those of non-IB students (Bishop, 1998; Gunderson, Maesch, & Rees, 1987). This favouring of one group over the other is described using directional values indicating which of the two groups has a higher chance of success on certain items. For example, if positive impact indicates females overall perform better than males on a test, then negative impact would indicate males outperform females on the test. In the context of this study, bias is attributed to any unfair characteristics of the item (e.g., wording), while impact is due to inherent differences that pre-exist between groups (Ackerman, 1992).

Therefore, a good test would detect this impact between the two groups without any biased items. However, the existence of impact and bias complicates DIF detection because they both contribute to systemic group differences in responses. Hence, when CATSIB detects a difference between two groups on an item, it is not able to differentiate whether this difference is due to the bias or impact. As such, there is a need to investigate the performance of CATSIB in detecting DIF items when impact is present. This study builds on a previous study that investigated the performance of CATSIB to detect DIF in a MST environment when: (a) item difficulty, (b) sample size, and (c) balanced/unbalanced design were manipulated (Gierl et al., 2013). However, Gierl and colleagues (2013) did not investigate the effects of CATSIB in detecting DIF when impact was present.

Purpose

The purpose of the present study is to evaluate the performance of CATSIB in detecting DIF in a multi-stage adaptive testing environment when impact is introduced to the system. The specific research questions addressed were:

1. Does CATSIB detect DIF items adequately, as measured using power and type I error (TIE) rates, when no impact is present?
2. How does the introduction of impact affect CATSIB's ability to detect DIF items adequately?
3. How does the direction of the impact, whether the focal group is performing higher or the reference group is performing higher, affect CATSIB's ability to detect DIF items adequately?

This research aims to inform test users and developers of potential issues with DIF items and add to the literature on DIF in a MST environment.

Method

Simulations

A MST environment was created using the R programming language (R Development Core Team, 2011) with specific levels impact introduced to the administration (Gierl et al., 2011). The R programming code used for this study is available upon request from the first author. The simulation consisted of 300 items administered to 7,000 students. The reason to use these numbers of items and students is to control for relative over exposure and robustness of each item. To ensure each item is not over exposed to students, a ratio of 1 item to 500 students is recommended from IRT research (Gierl et al., 2011). The adaptive nature of the MST environment created similar numbers of examinees in the focal and reference group. The simulated items within the items bank were made to mimic real life exam situations where a variety of easy, moderate, and difficult items are administered to examinees. The difficulty levels of each item, known as the b-parameter in IRT, were set so that easy (*M* = -1.25, *SD* = 0.50), moderate (*M* = -0.25, *SD* = 0.25), and difficult (*M* = 0.25, *SD* = 0.25) were equally represented in each module of the simulation (Gierl et al., 2011). Within the exam bank, 10% of the items were systematically programmed to contain large levels of DIF. Each simulated examinee 'wrote' seven modules of four items each indicating approximately 10% exposure rate for the exam bank. Due to the adaptive nature of the exam, approximately 250 to 275 examinees in each comparison group wrote each module of items, which is the minimum size recommended for adequate results when trying to detect biased items using CATSIB (Gierl et al., 2011). The simulation for each condition was replicated 100 times.

In total, 13 sets of simulated data were created to test CATSIB's performance under different impact settings. To address the first research question, where no performance difference was present, one set of simulations were created where both groups of examinees had equal and average abilities, centered on zero. This simulated environment is represented as "1. None" in Table 1. Results of this environment will create a baseline for comparisons with environments that contain impact. To address the second research question, setting the reference group performance to be average and the focal group abilities to be 0.5, 1.0, and 1.5 standard deviations above the focal group introduced performance differences. These mean abilities were

Table 1

The Power and Type I Error Rates for Various Impact Groups with Different Item Difficulties

Impact	Statistic	Total (%)	Easy (%)	Moderate (%)	Difficult (%)
1. None	Power	81^{a}	83^{a}	80^{a}	80^{a}
	TIE	5^{a}	4^{a}	6^{a}	5^{a}
2. Focal 0.5*	Power	85^{b}	80^{e}	86^{e}	89^{c}
	TIE	6^{b}	6	7	4^{c}
3. Focal 1.0	Power	82^{b}	72^{e}	84^{e}	91^{c}
	TIE	10^{b}	15	10	5^{c}
4. Focal 1.5	Power	77^{b}	66^{e}	76^{e}	90^{c}
	TIE	20^{b}	34^{d}	18	7^{c}
5. Reference 0.5	Power	74^{b}	74^{e}	74^{e}	74^{c}
	TIE	6^{b}	6	7	4^{c}
6. Reference 1.0	Power	69^{b}	67^{e}	71^{e}	68^{c}
	TIE	10^{b}	16	10	5^{c}
7. Reference 1.5	Power	66^{b}	67^{e}	69^{e}	63^{c}
	TIE	20^{b}	35^{d}	18	7^{c}
8. Focal -0.5	Power	85^{b}	87^{c}	86^{e}	83^{e}
	TIE	6^{b}	3^{c}	7	9
9. Focal -1.0	Power	84^{b}	90^{c}	84^{e}	79^{e}
	TIE	11^{b}	3^{c}	10	20
10. Focal -1.5	Power	79^{b}	88^{c}	77^{e}	72^{e}
	TIE	21^{b}	4^{c}	19	40^{d}
11. Reference -0.5	Power	76^{b}	79^{c}	75^{e}	74^{e}
	TIE	6^{b}	3^{c}	7	9
12. Reference -1.0	Power	70^{b}	73^{c}	71^{e}	66^{e}
	TIE	11^{b}	3^{c}	11	19
13. Reference -1.5	Power	67^{b}	72^{c}	67^{e}	62^{e}
	TIE	21^{b}	4^{c}	18	40^{d}

Note. *Indicates mean performance for focal group is 0.5 and reference group is 0 chosen because they have been previously compared in SIBTEST and provide a basis to compare the results of the current study (Gotzmann, Vandenberghe, & Gierl, 2000; Narayanan & Swaminathan, 1994; Roussos & Stout, 1996). Impact can favour the focal group, where the focal group's abilities are higher than the reference group, but it can also favour the reference group, where the reference group's abilities are higher. Hence, nine sets of data were created to test these two directional situations. Since the variable of DIF was not manipulated, it would be presumed that the simulated impact caused any changes in CATSIB results of impact environments.

Data Analysis

Two dependent variables were used to evaluate the performance of CATSIB in this study, power and TIE rates. Power is the probability of correctly detecting a DIF item and was calculated based on the percentage of DIF items that were correctly identified as having DIF (Gotzmann & Boughton, 2004). TIE refers to the probability of incorrectly identifying an item as DIF when it in fact does not, and was calculated using the percentage of times a non-DIF item was identified as being a DIF item (Gravetter & Wallnau, 2009). High TIE values indicate CATSIB calculated poor estimates, which usually occurs when extreme samples (e.g., small sample size) are present. Generally, a power rate of 80% and TIE of 5% indicates relatively good standards of detecting DIF items (Glass & Hopkins, 1996).

Item responses generated from the simulation were analyzed for DIF using CATSIB. Performance of CATSIB in detecting DIF was calculated using power and TIE rates. The independent variable of this study was the mean performance of each group of students (13 environments), difficulty of items (three difficulty levels), and which items each student wrote (adaptive nature of MST). The next section will describe the results of CATSIB.

Results

A summary of the results is presented in Table 1 and coded using superscripts. The superscripts correspond to different sections of the results presented. To address the first research question of no impact present between the two groups of examinees, as shown with cells denoted with ^{a} in Table 1, CATSIB detected DIF adequately with a power rate of 81% and TIE rate of 5%. These results are similar to the results of previous studies where SIBTEST investigated the effects of DIF items without impact (Gotzmann et al., 2000).

To address the second research question of impact being introduced to the system, results show that as the impact increased, the power rates decreased while the TIE increased. These results are represented in Table 1 with cells denoted with ^{b}. The inverse changes in power and TIE rates were similar regardless of positive or negative impact in the system. This means that as the abilities of the two groups differed, such as during comparisons between a non-IB and IB class, CATSIB was not able to adequately detect DIF items.

To address the third research question of directional impact being introduced to the system, three findings were seen. First, if the direction of impact and DIF is known we can minimize power loss due to the magnitude of impact. Second, in an adaptive testing environment, knowing the first conclusion, we can use CATSIB for certain aspects of information. For example, if we compared a group of regular (reference) and higher performing IB (focal) students we would focus on environment (2) through (4), particularly the difficult items because of the higher power and low TIE rates. We could use the CATSIB results shown in the cells denoted with [c], with confidence that DIF items are consistently being detected for the group of regular and IB students. Third, the cells denoted with [d] indicate extremely high TIE rates.

It is interesting to note that when the reference group's abilities were being manipulated to be higher or lower than average, none of the power or TIE rates were adequate, even when very little impact, such as environments (5) and (11), was present. This was evident in all the environments where the focal group's abilities were manipulated and for all levels of item difficulties. Our investigation of evaluating the performance of CATSIB in detecting DIF in a simulated MST environment guided these results. The next section will use the results from CATSIB to discuss the significant findings of this study.

Discussion

The results of this study showed that under a variety of environments, different levels of item difficulties and performance levels of examinees, the ranges of power and TIE rates differed. When no impact is present, the results indicated that there was adequate power and TIE rates. However, as impact was added to either the reference or focal group, the power decreased and TIE rates increased. Thus, in a MST environment, the amount of DIF present is not equal for all items when impact is introduced to the system. There is variability in how well CATSIB is able to detect biased items as the program does not favour all samples of examinees equally. That means when we know the type of students we are comparing and the item difficulty, we are able to look at the specific cells to determine which environment and difficulty levels do not have adequate power and TIE, such as the cells denoted with [e], and avoid those areas. On the flip side, we are also able to determine which areas, such as the cells denoted with [c], have adequate power and TIE and use those areas of results.

In some extreme cases, as shown in the cells denoted with [d], the inflated TIE rates could be attributed to the estimation calculations being made with fewer students in the sample. The low power and high TIE rates in these environments of a high performing student writing an easy item could be due to the adaptive nature of MST exams streaming fewer students towards these items and thus the estimated rates are calculated using fewer examinees. These extreme cases represent the statistical sensitivity of CATSIB in detecting DIF items, which is dependent on the sample size used to calculate the rates. Despite relatively weak power and TIE rates in extreme environments, the results are promising when the difficulty of the item and the nature of examinees' abilities are known. Thus, in an adaptive testing environment, when interpreting the DIF one should be aware of the fluctuation in the testing environment as well.

Another major finding that was consistent throughout out the different performance environments and item difficulty levels was that when the reference group's abilities were being manipulated, none of the power and TIE rates were adequate even when very little impact, such as environments (5) and (11), was present. This could be due to the fact that this simulation study contained only DIF items that favoured the focal group. Additionally, this phenomenon is an artefact of CATSIB, as well as other statistical methods of detecting DIF items, such as not being able to differentiate between biased items and impact between groups (Ackerman, 1992; Holland & Wainer, 1993; Nandakumar & Roussos, 2004).

This simulation study shows CATSIB performs relatively well when impact is introduced into the system when specific conditions are known. Even when extreme levels of impact are introduced, a difference of 1.5 between the two groups of examinees, a subset of the results can be used to investigate DIF items present in the exam. It is important to remember CATSIB is only a statistical tool used to help test developers detect DIF items. Thus, once the DIF items are detected, they are given to content specialists to verify whether they are indeed biased items.

Limitations and Future Studies

Two major limitations were identified in this investigation of power and TIE rates of CATSIB when impact was introduced. First, this study used simulated data to mimic real-life examinee responses, which tend to be slightly different from using real examinee responses from an actual exam (Gotzmann & Boughton, 2004; Gotzmann, Wright, & Rodden, 2006; Roussos & Stout, 1996). Second, this investigation used one variant of CAT assessments, namely a MST environment with three levels of item difficulty, which unavoidably incurred limitations as other variants of CAT exams are possible (Mean, 2006; Weiss, 1982). Therefore, future studies should use real examinee responses, with known performance characteristics, to evaluate the performance of CATSIB in detecting DIF items to provide more realistic and a richer data source. To fully investigate biased items in adaptive environments, future studies should consider testing other variants of CAT exams, such as a MST environment with more than three levels of item difficulty or a CAT environment where the test is adapted after every item administered to evaluate the performance of CATSIB.

It is surprising how little the area of DIF in CAT has been researched over the last decade despite the explosion of research on and application of CAT exams (Zwick, 2010; Gierl et al., 2011). As CAT exams, such as MST, become popular it is important to ensure these tests are developed so that they are fair to all examinees. The next sections present the conclusions and educational significance.

Conclusions

Assessments for the 21st century have a strong person approach where tests are tailored for individual students instead of the typical uniform exam for all students regardless of their performance levels (DiCerbo & Behrens, 2012; Shute & Ventura, 2013). CAT environments have been able to provide students with this level of individualized tests, which tailor items according to their performance levels. In addition to CAT exams, MST environments allow modules of items to be administered and adaptations of the test are made based on those performance estimates. Of course, with increased innovations in assessment, there should also be more research to assess whether these new forms of testing are viable in terms of meeting the basic principles for fair student assessments (Joint Advisory Committee, 1993). Thus, this study simulated a MST environment to mimic real world testing where students of various abilities may be presented with the same item at the beginning of the examination. The simulated results, with various levels of impact introduced to the system, were processed through CATSIB for DIF, or biased item, detection. The efficiency of CATSIB being able to detect DIF was measured using two dependent variables, namely the power and TIE rate, which are expected to be 80% and 5% respectively for relatively good results. In the cases where the impact favoured a particular group and the difficulty level also favoured the performance levels of that group, then CATSIB was able to detect DIF items relatively well. Thus, if the item difficulty and abilities of the two groups of examinees are known then sub-sections of the DIF results from this CATSIB simulation study will be very beneficial in detecting biased items.

Educational and Practical Significance

The results presented in this study can aid test developers in terms of which sub-sections of CATSIB results to use to create fair MST assessments when the groups compared have systematic differences in overall performance. This statistical process will decrease the financial strain because hiring groups of content experts to review items can be a costly process. Thus, from the use of the CATSIB program, items identified as DIF can be streamlined and the content specialists will be able to devote their time to reviewing a group of selected items. The final beneficiaries of these enhancements towards MST assessments are students and examinees who will be given fair and equitable adaptive assessments.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*(1), 67-91.

American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME]. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Institute of Certified Public Accountants. (2012, July 24). *Certified public accountant exam*. Retrieved from http://www.aicpa.org/becomeacpa/cpaexam/pages/cpaexam.aspx

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.

Bennett, R. E. (2006). Inexorable and inevitable: The continuing story of technology and assessment. In D. Bartram & R. K. Hambleton (Eds.), *Computer-based testing and the internet: Issues and advances* (pp. 201-217). West Sussex, England: John Wiley & Sons.

Bishop, J. H. (1998). The effect of curriculum-based external exit exam systems on student achievement. *The Journal of Economic Education, 29*(2), 171–182. doi:10.1080/00220489809597951

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.

Brennan, R. L. (2006). *Educational measurement.* Westport, CT: Praeger Publishers.

Camilli, G. (2006). Test fairness. In R. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 221-256). New York: American Council on Education & Praeger series on higher education.

Camilli, G., & Shepard, L.A. (1994). *Methods for identifying biased test items.* Thousand Oaks, CA: SAGE Publications.

Chang, H-H., & Ying, Z. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23*(3), 211-222. doi:10.1177/01466219922031338

Cizek, G. J. (2009). Reliability and validity of information about student achievement: Comparing large-scale and classroom testing contexts. *Theory Into Practice, 48*(1), 63-71. doi:10.1080/00405840802577627

DiCerbo, K. E. & Behrens, J. T. (2012). Implications of the digital ocean on current and future assessment. In R. Lissitz & H. Jiao (Eds.), *Computers and their impact on state assessment: Recent history and predictions for the future* (pp. 273-306). Charlotte, North Carolina: Information Age Publishing.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Downing, S. M., & Haladyna, T. M. (2006). *Handbook of test development*. Mahwah, NJ: Lawrence

Erlbaum Associates.

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.

Gershon, R. C. (2005). Computer adaptive testing. *Journal of Applied Measurement, 6*(1), 109-127.

Gierl, M. J., Lai, H., & Li, J. C-H. (2011). *Evaluating the performance of CATSIB in a multi-stage adaptive testing environment*. Retrieved from Medical Council of Canada website: http://mcc.ca/wp-content/uploads/Technical-Reports-Gierl-Lai-Li-2011.pdf

Gierl, M. J., Lai, H., & Li, J. C-H. (2013). Identifying differential item functioning in multi-stage computer adaptive testing. *Educational Research and Evaluation, 19*(2-3), 188-203.

Gierl, M. J., Rogers, W. T., & Klinger, D. (1999, April). *Consistency between statistical procedures and content reviews for identifying translation DIF.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, QC.

Glass, V. G., & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd ed.). Needham Heights, MA: Allyn & Bacon.

Gotzmann, A. J., & Boughton, K. A. (2004, April). *A comparison of type I error and power rates for the Mantel-Haenszel and SIBTEST procedures when the group differences are large and unbalanced.* Paper presented at the American Educational Research Association, San Diego, CA.

Gotzmann, A. J., Wright, K. R., & Rodden, L. (2006, April). *A comparison of power rates for items favouring the reference and focal group for the Mantel-Haenszel and SIBTEST procedures.* Paper presented at the American Educational Research Association, San Francisco, CA.

Gotzmann, A., Vandenberghe, C., & Gierl, M. (2000, May). *Differential item functioning on Alberta achievement tests: A comparison of SIBTEST and TestGraf using data from native and non-native students*. Poster presented at the Canadian Society for the Study of Education Conference, Edmonton, AB.

Gravetter, F. J., & Wallnau, L. B. (2009). *Statistics for the behavioral sciences*. Belmont, CA: Wadsworth.

Gunderson, C. W., Maesch, C., & Rees, J. W. (1987). The gifted/learning disabled student. *Gifted Child Quarterly, 31*(4), 158-160.

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

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Joint Advisory Committee, Centre for Research in Applied Measurement and Evaluation. (1993). *Principles for fair student assessment practices for education in Canada*. Retrieved from: http://www2.education.ualberta.ca/educ/psych/crame/files/eng_prin.pdf

Li, J. C.-H., Gierl, M., & Lai, H. (2013). *A modified CATSIB procedure for detecting differential item functioning on computerized adaptive tests*. Manuscript submitted for publication.

Lightstone, K., & Smith, S. M. (2009). University student choice of computer vs. traditional paper-and-pencil tests: What predicts preference and performance? *International Journal of Technologies in Higher Education, 6*(1), 30-45.

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

Medical Council of Canada. (2012, July 24). *Qualifying examination part I*. Retrieved from http://www.mcc.ca/en/exams/qe1/

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing, 19*(4), 477-496. doi:10.1191/0265532202lt241oa

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

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement, 18*, 315-328.

R Development Core Team. (2011). *R: A language and environment for statistical computing.* (Version 2.13.0.)[Computer software]. Vienne, Austria: R Foundation for Statistical Computing.

Roussos, L., & Stout, W. (1996). 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.

Scalise, K., & Gifford, B. R. (2006). Computer-based assessment in e-learning: A framework for constructing "intermediate constraint" questions and tasks for technology platforms. *Journal of Teaching, Learning and Assessment, 4*(6), 4-43.

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.

Shute, V. J., & Ventura, M. (2013). Stealth assessment: Measuring and supporting learning in video games. Cambridge, MA: The MIT Press. Retrieved from http://myweb.fsu.edu/vshute/pdf/Stealth_Assessment.pdf

Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of construct representation. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 329-347), Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Way, W. D., Twing, J. S., Camara, W., Sweeney, K., Lazer, S., & Mazzeo, J. (2010). *Some considerations related to the use of adaptive testing for the common core assessments.* Retrieved from http://images.pearsonassessments.com/images/tmrs/tmrs_rg/TMRS_WP_CAT_Paper_common_core_11.03.10.pdf?WT.mc_id=TMRS_Some_Considerations_Related_to_the_Use_of_Adaptive

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

Zenisky, A. L., & Sireci, S. G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education, 15*(4), 337-362.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores.* Ottawa ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

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, NY: Springer Science and Business Media.

Man-Wai Chu (manwai@ualberta.ca) is a Ph.D. Candidate at the Center for Research in Applied Measurement and Evaluation in the Department of Educational Psychology, University of Alberta. Her research interests focus on educational assessments, particularly the use of innovative technologies, cognitive advancements, and different measurement techniques to improve the measurement of performance-based skills.

Dr. Lai is Assistant Professor and Director, Evaluation and Assessment of the undergraduate medical education program at the University of Alberta. His research interests include education assessment, item generation, adaptive learning, and computer-based testing.

Correspondence pertaining to this article should be directed to Man-Wai Chu by airmail at 6-110 Education North, Centre for Research in Applied Measurement and Evaluation (CRAME), Dept. of Educational Psychology, Faculty of Education, University of Alberta, Edmonton, Alberta, CANADA T6G 2G5 or email at manwai@ualberta.ca.