Language Testing in Asia

REVIEW Open Access

# An overview of differential item functioning in multistage computer adaptive testing using three-parameter logistic item response theory

Karim Sadeghi[1*†] and Zainab Abolfazli Khonbi[2†]

* Correspondence:
k.sadeghi@urmia.ac.ir
†Equal contributors
[1]Department of English Language and Literature, Faculty of Humanities, Urmia University, Urmia 165, Iran
Full list of author information is available at the end of the article

## Abstract

As perfectly summarised by Ida Lawrence, "Testing is growing by leaps and bounds across the world. There is a realization that a nation's well-being depends crucially on the educational achievement of its population. Valid tests are an essential tool to evaluate a nation's educational standing and to implement efficacious educational reforms. Because tests consume time that otherwise could be devoted to instruction, it is important to devise tests that are efficient. Doing so requires a careful balancing of the contributions of technology, psychometrics, test design, and the learning sciences. Computer adaptive multistage testing (MSCAT) fits the bill extraordinarily well; unlike other forms of adaptive testing, it can be adapted to educational surveys and student testing. Research in this area will be an evidence that the methodologies and underlying technology that surround MSCAT have reached maturity and that there is a growing acceptance by the field of this type of test design" (from the Foreword to Y. Duanli, A. A. von Davier, & L. Charles (Eds.), *Computerized multistage testing: theory and application*). This state-of-the-art paper aims to present an overview of differential item functioning (DIF) in MSCAT using three-parameter logistic item response theory (IRT), offering suggestions to implement it in practice with a hope to motivate testing and assessment researchers and practitioners to initiate projects in this under-practiced area by helping them to better understand some of the relevant technical concepts.

**Keywords:** Computer adaptive testing, Differential item functioning, Item bias, Item response theory, Multistage computer adaptive testing

## Background

### Item response theory

Dating back to mid-twentieth century, a new theoretical basis for educational and psychological testing and measurement has emerged which has been called latent trait theory also known nowadays as item response theory (IRT). Though the understanding of the concepts and issues in IRT is somewhat hard for the novice tester, IRT-based research has attracted the attention of many researchers interested in measurement and testing for a number of reasons: (a) IRT has the possibility of comparing between the latent traits of individuals from a variety of populations when they are subjected to tests or questionnaires that have certain common items; (b) it also allows for the comparison of individuals of the same population submitted to totally different tests; this is possible because in IRT, the items—not the tests or the questionnaire as a whole—are regarded as central elements (Andrade et al. 2000); (c) it also allows for a better analysis of each item that makes up the measurement instrument since the scale-building characteristics specific to it are considered; (d) the items and the individuals are in the same scale so that the level of every single individual's characteristic can be compared to the level of the characteristic which the item demands. This indeed makes the interpretation of the resulting scale easier and allows them to know which items are producing information throughout the scale (Embretson and Reise 2000); (e) IRT allows for the treatment of a group whose data is missing through the given responses alone, an exercise which is impossible in CTT; and (f) it follows the principle of invariance which means that the item parameters do not depend on the respondent's latent traits and that the parameters of individuals are not dependent on the given items (Hambleton et al. 1991). Generally speaking, IRT presents mathematical models for the latent traits which in turn leads to proposing forms of representing the relation that exits between the probability of an individual responding correctly to a given item, his latent trait, and characteristics (parameters) of the items (Andrade et al. 2000).

In order to be better able to understand the concepts used in the IRT literature, a brief overview of them is presented next. Technical words used in IRT estimates consist of severity, measure, theta, person, and item calibration location which are different on a scale from 0 to 100 and whose function is dependent on who and what proficiency level the test taker is. There are also item discrimination (the degree to which each item distinguishes person abilities), item information (an item is so specifically designed that it would be appropriate for information we want to derive about one's ability based on that item), and test information (the sum of item information or how the test discriminates ability levels). Moreover, some IRT statistics are parameters featuring every item, including item difficulty (the difficulty level of the item), item discrimination (how well the item discriminates people at a certain ability level on an item with a certain difficulty level), item parameter for guessing/pseudo-chance (somehow in the case of multiple-choice items related to differential item functioning and whether the item performs differently for different individuals). So all in all, what IRT offers is to consider each person's score with reference to the group it belongs to, i.e., estimates of the ability level are interpreted taking into account other factors like age, gender, and cultural background.

### *Three-parameter logistic item response theory*

There are three existing IRT models that are distinguished from each other in the mathematical form of the item's characteristic function and/or the number of parameters which are determined by the model. All the models can contain one or more parameters related to the items and the individual (Andrade and Bortolotti 2007). The main distinction between IRT models lies in the assumption about the relationship made between the options in the answer of a question and the level of the latent trait. Dealing with these three IRT models lies beyond the scope of the present discussion; however, in this overview paper, we have focused on the three-parameter logistic (3PL) IRT model, which is briefly described below.

The three-parameter logistic item response theory (3PL IRT) model is used for data coming from multiple-choice test items where an individual is expected to select from among the given alternatives the option that is the correct answer; however, there is also a probability that an individual may end up with the correct response simply by chance. That is, in addition to the two parameters of discrimination and difficulty, there is also a parameter of "guessing" shown as $\eta$ in 3PL IRT. So the formula for estimating the probability of an individual's correct answer on an item becomes (for further detailed explanations, see Curtis 2010)

$$pij = P(Yij = 1|\theta i, \alpha j, \delta j) = \eta j + (1 - \eta j)\frac{1}{1 + \exp\{-\alpha j(\theta i - \delta j)\}}$$

Here, $P_{ij}$ is the probability of a randomly chosen examinee with ability $\theta$ (measured in logits) answering item $j$ correctly the response is itself a random variable $Y_{ij}$ with a Bernoulli distribution (a special case of binomial distribution, McCullagh and Nelder 1989). $k$ is the number of items, $\alpha$ is called the discrimination parameter for item $j$, and $\delta_j$ is the difficulty parameter of the item measured in logits.

### *IRT and DIF*

During the last several decades, measurement specialists, substantive researchers, and the general public have become increasingly concerned with the possibility that psychological measures may "work differently" or be biased either for or against certain groups of examinees (e.g., men or women). For this reason, a rather large research base has accumulated regarding how to scrutinize psychological measures for item bias. According to Drasgow (1987), in any discussion of how a psychological measure functions across various examinee groups, a distinction should be made between two types of bias: external and internal. External bias happens when test scores have multiple correlations with non-test variables for two or more groups of examinees. This results in differential predictive validity of the measure. This differential validity or lack of structural invariance may or may not be of concern depending on the context of test use. In many research contexts, the differential predictiveness of an estimate is measured by a substantive theory and may form the heart of a research program. The second form of bias is when the internal relations of a test (e.g., the covariances among item responses) differ across two or more groups of examinees. This is called *measurement bias* which in turn leads to a measurement scale not being invariant or equivalent across groups. The study of a scale's item-level measurement invariance is of fundamental importance in psychometric research. In order for test scores (either observed raw scale scores or estimates of the levels of the latent trait) to be comparable across various groups of examinees, those scores must be on the same measurement scale. That is, the items must display measurement invariance across different groups of examinees.

In more technical terms, differential item functioning (DIF) (which replaces bias in IRT) is said to occur when a test item does not have the same relationship to a latent variable (or a multidimensional latent vector) for two or more groups of examinees. Obviously, an item displays DIF if the item characteristic curve differs for different groups or, similarly, when each/any of the item parameters differs across groups. In other words, DIF is present when individuals from different examinee groups have different likelihoods or probability of answering an item correctly, after controlling for ability. Controlling for group differences in ability is a critical step in any DIF analysis, because it matches examinees on a common measure before comparing them.

### Computer adaptive testing

In line with the developments in the field of foreign or second language teaching where computers are being highly favored to assist language teaching and learning, the field of language assessment is also increasingly making use of computers. One of the chief potential applications of IRT models lies in the realm of computerized adaptive testing (CAT). A CAT is a measurement instrument that uses computers in administering tests to examinees. Perhaps, the major goal of CAT is to apply the IRT invariance property to develop an algorithm by which a test is given to an individual that is a "good" measure for that particular examinee. So it can be claimed that the test is "tailored" to each examinee so that the questions are neither too difficult nor too easy. As a by-product, IRT-based CAT exams usually contain fewer items than conventional paper-and-pencil measures. CAT algorithms based on IRT have been the subject of various investigations for several decades, and much is known about their functioning in comparison to paper-and-pencil measures (Wainer et al. 1990).

Similarly, more than a decade ago, Bennett (2001) claimed that in large-scale assessments, no topic would become more subjected to innovation and future practice than computers. His prediction has proven to be accurate. Large-scale assessment and computer technology have developed staggeringly since 2001. Consequently, many large-scale assessments, which were once administered in a paper-and-pencil format, are now administered with the help of computers. Computerized assessments bring many advantages to examinees compared to more conventional paper-based assessments. For instance, computers support the creation of new alternative item types and innovative item formats (Sireci and Zenisky 2006); items on computer-based tests can be scored immediately to offer immediate feedback to the examinees (Drasgow and Mattern 2006); and computers also allow on-demand testing (van der Linden and Glas 2010). However, the most significant benefit of computerized assessment is that it permits examiners to assess more complex performances of students via integrative test items which may include digital media to increase the types of skills, knowledge, and competencies that can be evaluated (Bartram 2006).

#### *CAT and DIF*

Conducting DIF analyses with CAT becomes of utmost significance to ensure that examinees are tested without bias. Because testing has become a global enterprise, heterogeneous samples of examinees with different cultures, languages, learning opportunities, educational backgrounds, skills, knowledge, and access to technology and computers are involved in the same exams that are expected to yield the same score interpretations. For instance, Phillippe Grosskost, the Managing Director of ETS Global for Europe, Africa, and the Middle East, asserted 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, p. 4). This strong claim highlights the importance of testing without bias. In addition, the need for DIF-free item administration also stems from the adaptive nature of CAT. As Zwick (2010) noted, examinees write fewer items in an adaptive testing context, which implies that every item has more contribution to the final estimate of the ability. So the presence of item bias could exert a stronger effect on the ability estimates of the examinees. Bias could also influence the order of item administration, assuming that the selection of items on an adaptive test is specified by the candidates' responses to the items presented earlier such that examinees with better performances on previous items will receive a more difficult item whereas examinees with poorer performances will receive a less difficult item (Chang et al. 2011).

However, what is particularly challenging when conducting DIF analyses with CAT is that CAT requires large numbers of items since item banks are required to allow continuous testing while simultaneously limiting the amount of exposure to the item. As a result, these large banks of items must first be developed and continually restored in order to reduce item exposure and keep test security while allowing for continuous administration of the test. At the same time, procedures, policies, and reviews must be designed and applied to ensure that items meet the basic standards anticipated for equity and fairness. In this regard, sensitivity reviews are conducted to ensure that items used for CAT observe these basic standards (Educational Testing Service 2007). Since panelists can focus their interpretations of test scores on those items which entail large DIF estimates and, thereby, those which produce large differences in groups, sensitivity reviews are then also informed by the outcomes from such analyses of DIF. However, unfortunately, relative to the items on a paper-based exam, the number of examinees on an adaptive test who answer any one item may be small, especially when there is a large item bank. Accordingly, DIF methods designed to help monitor fairness in CAT must function in diverse testing environments and, often, when there are large number of items in the bank but the number of individual examinees who respond to each one of these items is rather small.

Traditionally, there were programs developed to conduct DIF on paper-and-pencil tests, for example, programs such as SIBTEST (Shealy and Stout 1993) that used a transformation that accounts for differences in the reliability of total test score as a conditioning variable to estimate an examinee's true score. This is possible because all items are administered to all examinees. However, with the use of CAT becoming increasingly popular, new techniques were needed that would permit testing for DIF in case not all examinees answer the same items or even the same number of items. Specifically, an alternative conditioning variable was needed in this type of testing situation (Walker et al. 2001). The program developed for DIF analysis in CAT was called Computer Adaptive Test-Simultaneous Item Bias (CATSIB) (Roussos 1996) that allowed identifying DIF items through estimating an examinee's ability for the conditioning variable, rather than a total test score. In order to control for estimation bias, and inflated type I errors (which are produced when the focal and reference groups vary in their distributions of observed score), CATSIB automatically uses what is naturally called regression correction.

Roussos (1996) developed CATSIB a modified version of SIBTEST that uses an examinee's ability estimate as the conditioning variable. In the conventional use of SIBTEST, observed score was used as an estimate of true score. At each estimated true score level, the proportion of examinees in both the focal and reference groups is achieved who obtained the correct response to the item being studied. Then, of the differences of these proportions for the two groups at each estimated true score level, a weighted average is calculated to obtain the test statistic, $\beta$. However, when using an examinee's estimate of ability as the conditioning variable, it is very unlikely that two examinees have the exact same value of $\theta$ since it is a real numbered variable, not an integer variable. So the observed range of $\theta$ is divided into $n$ intervals that are equal provided that at least from each group, three examinees are included in the interval (Nandakumar and Roussos 1997).

Therefore, by matching examinees on the conditioning variable and then summarizing the differences in performances over these levels, DIF methods correct for differences in true mean target ability between the focal and reference groups. However, with no more correction, the existence of impact still inflates DIF detection rates when none exists (Shealy and Stout 1993). With the presence of impact, even when there is no DIF, the expected value of the reference group's target ability will tend to differ from the corresponding expected value for the focal group: $E_R\{\theta|\,"\,\theta\} \neq E_F\{\theta|\,"\,\theta\}$. Thus, with no further correction, the expected value of the DIF test statistic will tend to be different from zero, causing an inflated type I error rate.

Some DIF detection procedures, such as the standardization procedure (Dorans and Kulick 1986) and Mantel–Haenszel (Holland and Thayer 1988), include the examinee's score on the studied item of concern (the one for which the DIF detection procedures are being used) in the total test score. The inclusion of the studied item facilitates controlling for type I error in the DIF detection for these procedures (Jiang and Stout 1998). However, SIBTEST uses a regression correction that was designed to correct this inflated type I error rate originating in the estimation bias of the statistic's ($\beta$). Shealy and Stout (1993) proposed that SIBTEST "allows the practitioner to … estimate accurately the amount of bias/DIF without contamination from target ability distributional differences across group" (p. 161). Similarly, CATSIB employs a regression correction to compensate for the potential estimation bias. Specifically, CATSIB matches examinees on an estimate of where the subscript $G$ represents which group an examinee comes from, rather than matching examinees on observed test score as SIBTEST does.

## MSCAT and DIF

### Need for multistage computer adaptive testing research

As mentioned earlier, computer-based testing has dramatically changed educational measurement research. Another type of preconstructed adaptive test that is gaining in popularity is computerized adaptive multistage test (MSCAT). An MSCAT has the same purpose as a CAT but uses larger units to build the test. Considering the lack of this type of research in the field of Teaching English as a Foreign or Second Language (TEFL/TESOL) and that MSCAT is not yet appropriately practiced for measurement purposes in such educational contexts, this brief overview of this measurement procedure is hoped to pave the way for better understanding of the issue and helping both testing bodies and researchers to consider the empirical findings of this type of testing to be better able to apply MSCAT in their contexts in hopes for more efficient assessments.

This expectation is in line with MSCAT's growth in popularity and implementation in well-known testing programs such as the Uniform CPA Examination of the American Institute of Certified Public Accountants which is used to license public accountants in the USA. It is also employed in the Qualifying Exam Part I of Medical Council of Canada, which is applied to admit medical students into supervised clinical practice. Not only DIF on CAT can negatively affect the adaptive item administration procedures but also the consequences of DIF on MSCAT could exert an adverse effect on the estimate of ability of the examinees since bias could influence the order of administration for the item sets on an MSCAT. To date, however, there has been little research on DIF in MSCAT; for an early evaluation of DIF in MST using the Mantel–Haenszel method, see Zwick and Thayer 2002).

Although the assessment of DIF is necessary in many assessment contexts, the significance of DIF is even greater when the examinees are compared based on their answers to different items, such as in computerized adaptive testing (Zwick 2010). This is the case because respondents are administered different items; then, the presence of DIF items can produce bias in a CAT situation both within a group and between the groups. The additional effect within groups happens since not all of the examinees are administered the items that exhibit DIF. So some of the respondents are advantaged and others are disadvantaged. In addition, fewer items are typically administered in a CAT. A DIF item can correspondingly have a large impact on the result of a test. An item that exhibits DIF can also have large repercussions in a CAT since the sequence of administered items to the respondents in part is dependent on their responses to that item (Makransky and Glas 2013).

Indeed in many cases according to Zumbo (1999), because of the fact that the test items contain sources of difficulty that are extraneous or irrelevant to the measured construct, the items in a test are biased. Perhaps, the item is tapping a secondary factor or factors over-and-above the one of interest. The subject of test bias has been the topic of a wide range of recent research. However, implementing 3PL IRT model to discover possible sources of DIF in an MSCAT context is a topic which is not yet adequately practiced in TEFL/TESOL educational contexts. Furthermore, physical security of test items is essential to ensuring that test-score interpretations are valid. While this security can be established independently from the testing mode, there are some benefits to the MSCAT format when security is considered. Since MSCAT makes use of interchangeable modules that can be shuffled into various forms, it may be easier to manage lower item exposure for smaller banks of test questions. This also decreases the likelihood of adjacent test takers seeing the same test form. Finally, stakeholders express a strong preference for giving candidates the capability to review and revise answers to test questions in the administration, MSCAT can be modeled to allow test takers to only review items within modules, and the opportunity is restricted for reviews between subtests.

Roussos and Stout (1996) developed a framework for DIF analysis to unify the substantive and statistical analyses that serves as one of the very first model-based approaches for identifying and interpreting the factors that elicit group differences. The independent variables that affect MSCAT's statistical performance are becoming more apparent, therefore researchers and practitioners can also apply this psychometric procedure to begin to evaluate the variables that could explain the presence of DIF. For instance, Zwick (2010) believed that test administrations based on computer might elicit several new and important sources of DIF that do not exist in paper-and-pencil testing, factors such as anxiety, differential computer familiarity, and facility. Hypotheses of these types can now be assessed, at least in some CAT situations, by carrying out MSCAT with the DIF analysis paradigm to study the factors that could explain why DIF occurs.

MSCAT is a sophisticated approach for delivering survey tests, and other measurements which is based on complex computer algorithms that through also controlling for practical issues, such as item exposure, test length, and content distribution, adapt the test to every examinee. Research shows that without a loss of precision, CATs reduce test length by up to 90%. However, to achieve such a reduction and make use of the advantages of CAT, it is inevitably essential that the developers of CAT carry out research investigations to simulate CAT performance. To date, however, MSTGen is the only software program that completely fills this role (Han 2013). All the points mentioned above signify the new developments and innovations in the field of CAT. Another important aspect of this overview note on this type of research is the use of three-parameter logistic item response theory to the identification of DIF in MSCAT. In TEFL/TESOL educational contexts, the idea of 3PL IRT-based DIF analyses as well as MSCAT are not satisfactorily examined. It is hoped that the advantages of the elements of these investigations (i.e., 3PL IRT, DIF, and MSCAT) tackled in once place will prove inspiring and insightful for prospective researchers and practitioners in the field of language assessment.

## Brief overview of DIF in MSCAT studies

MSCAT is not a new idea. A kind of non-computerized MST was developed prior to computerized adaptive tests (CAT; Lord 1971, 1980, chap. 9). However, MST research was eclipsed by CAT: the use of item response theory and computers in a creative manner offered the potential to design shorter tests than their predecessors which also enjoyed greater reliability estimates. Nowadays, various types of item-level adaptive tests have widespread uses, and CAT fully met its potential to create more efficient assessments. Nonetheless, as CATs began to be used in practice, certain practical shortcomings became evident, and a modern variety of MST arose to address these issues (Mead 2006). Multistage testing is an algorithm-based approach to administering tests which is very similar to CAT in that test items are interactively chosen for the examinees by the algorithm, but rather than selecting individual items, groups of items are selected, building the test in stages.

In their study on multistage adaptive testing for a large-scale classification test (design heuristic assembly, and comparison with other testing modes), Zheng et al. (2012) designed an MSCAT for a large-scale classification test and performed the automated test assembly using a heuristic method. Comparing then the performance of MST with that of a CAT through computer simulation and a linear form of test, they found that the automated assembly of tests was successful. The researchers also reported observing a trade-off in item bank usage and measurement accuracy when comparing MST and CAT. However, for classification purposes, MST proved enjoying as a good classification accuracy as that provided by CAT, with more efficient item bank usage. In a comparison of multistage tests with computerized adaptive and paper-and-pencil tests, Rotou et al. (2007) investigated the measurement precision of MST to CAT and paper-and-pencil tests for the three IRT models when the test is entirely set-based. The findings revealed that MST performed better in terms of reliability and conditional standard error of measurement for the 2- and 3-PL models than the same length paper-and-pencil test. Finally, results showed that MST performed better for the

1- and 2-PL models than the CAT test of equivalent length; however, MST and CAT performed approximately the same for the 3-PL model.

Recently in an investigation on designing and implementing a Multistage Adaptive Test: The Uniform CPA Examination, Melican et al. (2010) argued that the Certified Public Accountants (CPA) Examination was launched with the multiple-choice portions presented as an MSCAT. The criteria for choosing this model were primarily the ability to capitalize on the benefits of adaptive testing without limiting the ability of the test takers to revisit and change answers to previous items within subtests. The authors believed that the format may also allow future changes to introduce better, more efficient diagnostic information for the test taker. They reviewed the decision-making process from conceptualization of the new CPA Examination through the setting of the pass–fail standards to score reporting. They stated that the model had performed well since launch in 2004. They further added that panels and subtests met stringent validity and comparability standards, and the resultant scores were consistent across panels, routes, examinees, and administration windows. Melican et al. (2010) also highlighted that the computerized administration, assembly, and scoring model for MSCAT had been entirely accepted by the test takers. The process of MSCAT is intense, disciplined, informed by research and best practices, and completely transparent.

In another study on evaluating the content validity of multistage-adaptive tests, Crotts et al. (2012) evaluated content validity of an MSCAT using subject matter experts' content validity ratings of all items in the MSCAT bank. Analyses of these ratings across the most common "paths" taken by examinees were conducted. Their results indicated the content validity of the different exams taken by examinees (28 different exams totally) were roughly equivalent. Their analyses illustrate the types of investigations that could be done to evaluate test content within an adaptive context. Specifically focusing on the degree to which the items measured their intended cognitive and content areas, Kaira and Sireci (2010) examined the content validity of a multistage-adaptive test (the Massachusetts Adult Proficiency Test in Math). In their study, experts in the subject matter reviewed the items and the test specifications and developed two ratings for every item. These ratings included selecting the item's measured content area and cognitive level. Even if consistency was across the panels and paths in the MST with some exceptions, the researchers recommended that future investigations should "investigate more comprehensive measures of item quality, such as the degree to which the item measures its intended benchmark along an ordinal rating scale" (p. 23), rather than simply judging which cognitive levels and content areas should be measured by each item.

Gierl et al. (2011) conducted a study on evaluating the performance of CATSIB for detecting DIF provided items in the matching and studied subtest are administered in a realistic adaptive MST context. They concluded that for the large reference/moderate focal group and the large reference/large focal group conditions, CATSIB met the acceptable criteria of 80% power rate and 5% type I error rate. Likewise, Chu et al. (2012) found that CATSIB is well able to detect DIF in an MST environment where impact is introduced into the system with power rate of 81% and type I error of 5%. Their study also showed that as the ability level increases, the power rate decreases while type I error increases. Furthermore, the researchers stated if the direction of impact and DIF is known, the power loss resulting from the magnitude of impact can be minimized.

More recently, a whole book has been dedicated to discussing computerized multistage testing, its theory, and applications (Yan et al. 2014) wherein a comprehensive account is made of the basic issues of test design, item banking, and maintenance to different types of test assembly, to kinds of multistage designs including their routing, scoring, and equating, to more general detailed considerations of enhancing test reliability, validity, fairness, and security through various procedures, and finally to its wide range of applications and implementations in large-scale assessments. With all the advancements in the field of MSCAT globally and taking into account the wide range of topics that can be investigated in this area, still, the amount of research done on MSCAT is meager in the area of language testing. In order to widespread the idea of unbiased and fair test items in an adaptive manner, this paper aims to make researchers and practitioners in the field of language testing familiar with some of the advantages of and issues in MSCAT with an eye to make the practice of assessment more efficient and accurate.

### Multistage computer adaptive testing

The concept of multistage testing was introduced several decades ago (Lord 1971). However, the operational complexities of building modules to optimize measurement precision, balance content, and control exposure (especially over extended periods of time) were not anticipated (Luecht 2013). Luecht and Nungester (1998) provided one of the first large-scale practical design specifications for implementing MSCAT. Since that early design work, the MSCAT framework has been implemented for the Uniform CPA Examination (American Institutes of Certified Public Accountants) (see Melican et al. 2010) and several state examination programs. MSCAT is also being seriously considered as a replacement for item-level CAT for other high-stake large-scale examinations (e.g., the 2011 GRE). MSCAT is a balanced compromise between linear test forms (i.e., paper-and-pencil testing and computer-based testing) and conventional item-level CAT. It combines the advantages of both. On the one hand, MST is adaptive (so more efficient compared to linear tests). On the other hand, unlike CAT, it permits test developers to review different forms of test prior to administration, and it also permits examinees to review and revise their answers (Luecht 2013).

In this sense, MSCAT designs are structured adaptive tests that employ preassembled subtests as the basic units of test administration (Luecht and Nungester 2000). In contrast to item-level CAT designs, which result in different test forms for each test taker, MSCAT designs use a modularized configuration of preconstructed subtests and embedded score-routing schemes to prepackage validated test forms.

The idea of set-based tests with mechanical branching rules independent of IRT, administered via paper and pencil, can be traced to studies by Linn et al. (1969), Cronbach and Gleser (1965), and Angoff and Huddleston (1958), among others. When Lord and Novick (1968) outlined the fundamental assumptions of modern IRT, Lord (1971) was the first researcher to provide the framework and measurement justification for adaptive by-stage testing with IRT. Two-stage testing was described there as a method of obtaining improved measurement for not only typical examinees but also, and most importantly, those at the extreme ends of ability distribution. Nowadays, investigation into the alternative test designs within the broad heading of MSCAT in many contexts and domains is ongoing (Mead, 2006).

## The recommended procedure

The present overview introduces the procedure that can be used to assess the performance of CATSIB for detecting DIF on for example a researcher-made validated and reliable English proficiency test when items in the simulated study would be paper-based and in the real study would be administered in the context of a multistage computer adaptive test (MSCAT). The particular MSCAT could be simulated using a four-item module in a seven panel administration. The variables which would be expected to affect DIF detection rates could be assumed to be item difficulty level; participants' age, gender, language, educational, and cultural background; and computer and technology familiarity. It would be expected that using the three-parameter logistic IRT (difficulty, discrimination, and guessing), CATSIB could accurately and consistently detect DIF on an MSCAT of English proficiency among different focal and reference groups.

The simulated study, however, can consist of nearly 100 students the same as those in the sample of the real study. The participants may first take a paper-and-pencil test of English proficiency in order to test their homogeneity at the initial stage and match them against their English ability (matching subtest). The students in this group can then take part in an attitude toward paper-and-pencil assessment survey. In the real study, the students should sit in front of their computers with the CATSIB program already implemented on each computer (to ensure test security), and in the registration stage, they may fill out a background questionnaire form like asking for their age, gender, language, educational, and cultural backgrounds, as well as computer and technology familiarity. They can then start the MSCAT of their English proficiency. Finally, an attitude toward MSCAT questionnaire can be administered among the participants in order to see how they felt about their MSCAT experiences and to possibly find out their feedback (suggestions, weaknesses, etc.). So, first, the examinees can be given a paper-and-pencil test of English proficiency in order to control for the effect of proficiency affecting DIF. Then, in the next step in the construction of the computer-adaptive test, the researchers can estimate the difficulty of a large number of items that are to be used in the test. This process, known as item calibration, involves administering test items to a sample of examinees representative of the sample who will be given the computer-adaptive test.

Many different variations of MSCAT are permissible (Zenisky et al. 2010). However, some key concepts do exist across these variations. A block or set of items is referred to as a module (Luecht and 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 are 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 or 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.

To this end, an MSCAT environment can be developed using the R programming language (R Development Core Team 2005). The simulated MSCAT uses a four-item module in a seven-panel administration, where the first panel contains a single module that is common to all examinees, while the second through the seventh panels each contains three modules with items at three different difficulty levels. Using this structure, each examinee completes seven modules and writes a total of 28 items. Each module contains items at three levels of difficulty—easy, medium, and hard. The items can be generated using a 3PL IRT model. Figure 1 that follows clearly represents the model that can be followed in the MSCAT of English proficiency.

Examinees can be routed in the MSCAT using a simple strategy determined by their number correct score. Examinees who answer all four items correctly move to a harder module; examinees who score 0 or 1 move to an easier module; examinees who score 2 or 3 move to a module with the same difficulty level. The item bank size for the simulation stage can be fixed and contain 100 items at each difficulty level.

The time the study will last cannot be determined at this stage since in every administration as is obvious in all CAT, the number of examinees who write any one item is limited because item exposure rates should often be kept to a minimum. As a result, the DIF method designed to help monitor bias in CAT will be functioning in a testing environment where 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. For DIF studies using CAT administration in general, because the examinee item responses are dynamic, the expected level of DIF is assumed to be affected. That is, the real study examinee item responses produced with an adaptive testing process will supposedly affect the magnitude of DIF. To address this problem, the DIF items are to be

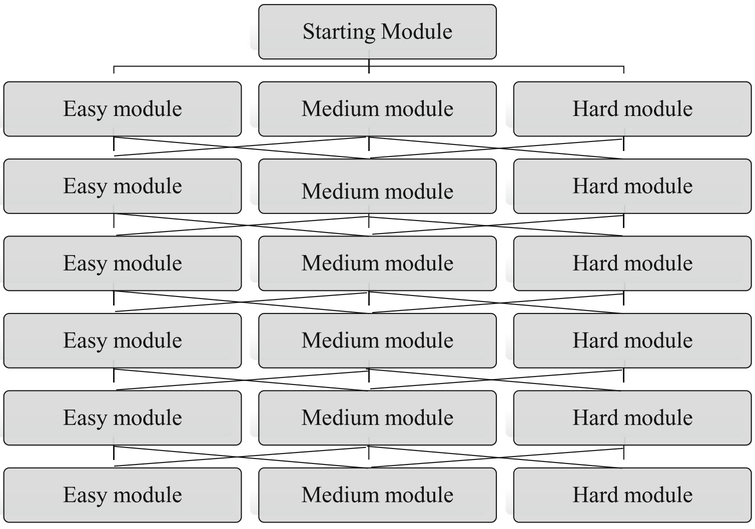

**Fig. 1** The model of MSCAT in the present study

generated according to a normal distribution, with a certain mean and standard deviation of " $\beta_{UNI}$.

### The rationale behind MSCAT

The concept of MSCAT was first proposed by Cronbach and Gleser (1957) in a personnel selection context for classification rather than measurement purposes. However Angoff and Huddleston (1958) at Educational Testing Service (ETS) first carried out MSCAT in an educational measurement context, in particular, in college admission testing. The rationale for using MSCAT given by Angoff and Huddleston was the changing needs of the admission testing process, in which the range of student skills was becoming increasingly broader, while at the same time, the same admission tests were beginning to be used in scholarship programs that demanded top students. In other words, high precision of measurement was required throughout the score scale. The two-stage design they developed was described by Angoff and Huddleston (1958, p. 1) as consisting of "two tests, each relatively homogeneous with respect to the distribution of item difficulties and pitched at different but overlapping levels of talent so that, in combination, they would embrace a wider range than the tests in operational use at the time."

Moreover, the more general idea of adaptive (Weiss and Betz 1973) or tailored testing (Lord, 1980) was fairly well elaborated, as noted by Wood (1973), by the early 1950s. Hick (1951), for example, noted that in theory, an intelligence test acts in a branch-model process: every individual who takes it at the first stage should have a 50% chance of being right and those who have been successful again should have 50% chance of answering the next items correctly and so on.

Bock and Mislevy (1988) went against the current at the time and argued that an MSCAT design they called the *duplex design* presented multiple advantages in school-based testing. The MSCAT has some efficiency afforded by tailoring to the ability level of the examinee and yields more precise scores, given fixed test lengths. However, this efficiency is not as great as CAT administration where adaptation occurs for each test item. However, Rotou et al. (2007) compared a two-stage multistage test to a CAT of the same length, both of which including only set-based items. The MSCAT had slightly higher reliability under the one- and two-parameter models and equal reliability under the three-parameter model. Also, given a fixed item bank size, due to the possibility of building a variety of different forms that overlap at the module level, MSCAT allowed the test developer to create a greater diversity of forms.

### Applications of MSCAT

The major advantage of CAT, relative to traditional paper-and-pencil exams, is that it is efficient, and research has repeatedly demonstrated that on average, an MSCAT exam is 50% shorter than a paper-and-pencil measure with equal or better measurement precision allowing researchers more time to assess the relevant constructs. Beyond this main feature, MSCAT offers the researcher other advantages as well. For example, the computer can continuously monitor the psychometric properties of the items; the examiner can easily record reaction time if this is of interest; and the computer administration opens up a wide array of alternative item formats to be included in the same exam. Furthermore, the computer can immediately score examinees and save that score to a database, and there is no need to buy answer sheets or number 2 pencils. There is a need to hire someone to constantly supervise the testing station, however. With networking and the rapid growth of the Internet, it is inevitable that a new world may be opened up to assessment researchers by computerized testing. As a result of these important advantages, many large-scale testing programs apply MSCAT, including the Graduate Management Admission Test (GMAT), the Graduate Record Exam (GRE), the Certified Public Accountants Licensure Exam, and the Armed Services Vocational Aptitude Battery (ASVAB). In short, an MST approach to adaptive testing is relevant to measure individual students more accurately and efficiently and to provide policy- and aggregate-level results to school administrators. Nevertheless, MSCAT, despite its virtues, has not been used extensively in education (Bejar 2014).

## Conclusions

This overview paper focused on introducing the process of identifying differential item functioning in the case of multistage computer adaptive testing following three-parameter logistic item response theory. It discussed some of fundamental concepts and considerations in these three areas and dealt with the rationale behind MSCAT along with a few studies done on it. Some applications of it and the need for research on this topic were also highlighted as well as a brief explanation on a proposed model for carrying out such research. The MSCAT research should be useful to psychometricians, researchers, and all those interested in latent trait theory and computer adaptive testing as technology advances. This type of research is intended for students, faculty researchers, practitioners at testing institutions, and education officers. It is hoped that readers of this introduction will find inspiration from the results of these kinds of studies and will approach the field of MST and, more generally, of adaptive testing with curiosity and interest in continuing the research presently under way and in making improvements to operational practice.

### Abbreviations

3PL IRT: Three-parameter logistic item response theory; CAT: Computer adaptive testing; DIF: Differential item functioning; IRT: Item response theory; MSCAT: Multistage computer adaptive testing

### Authors' contributions

Both authors read and approved the final manuscript.

### Competing interests

The authors declare that they have no competing interests.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Author details**
[1]Department of English Language and Literature, Faculty of Humanities, Urmia University, Urmia 165, Iran.
[2]Department of English Language and Literature, Faculty of Humanities, Kosar University of Bojnord, Bojnord, Iran.

Received: 28 December 2016 Accepted: 18 April 2017
Published online: 12 May 2017

### References

Andrade, D. F., Tavares, H. R., & Valle, R. C. (2000). *Item response theory: concepts and applications*. São Paulo: Associação Brasileira de Estatística (in Portuguese).

Andrade, D. F., & Bortolotti, S. L. V. (2007). Aplicação de um Modelo de Desdobramento Graduado Generalizado- GGUM da Teoria da Resposta ao Item. (Applying a Model Unfolding A Graduate Generalizado- GMM Theory of Item Response). *Estudos em Avaliação Educacional., 18*(37), 157–87.

Angoff, W., & Huddleston, E. (1958). *The multi-level experiment: a study of a two-level testing system for the College Board Scholastic Aptitude Test (statistical report no. SR-58-21)*. Princeton, NJ: Educational Testing Service.

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.

Bejar, I. I. (2014). Past and future of multistage testing in educational reform. In Y. Duanli, A. A. von Davier, & L. Charles (Eds.), *Computerized multistage testing: theory and application* (pp. 423–438). Boca Raton: Chapman and Hall/CRC.

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

Bock, R. D., & Mislevy, R. J. (1988). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431–444.

Chang, S.-R., Plake, B. S., Kramer, G. A., & Lien, S.-M. (2011). Development and application of detection indices for measuring guessing behaviors and test-taking effort in computerized adaptive testing. *Educational and Psychological Measurement, 71*, 437–459.

Chu, M. W., Lai, H., & Wang, X. (2012). Detecting directional DIF using CATSIB with impact present. *Paper presented at the Annual Meeting of the National Council on Measurement in Education,* Center for Research in Applied Measurement and Evaluation, University of Alberta, Vancouver, BC.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana, IL: University of Illinois Press.

Cronbach, L. J., & Gleser, G. C. (1957). Psychological tests and personal decisions. Champaign: University of Illinois Press.

Crotts, K., Sireci, S. G., & Zenisky, A. (2012). Evaluating the content validity of multistage-adaptive tests. *Journal of Applied Testing Technology, 13*(1), 1–26.

Curtis, S. M. (2010). BUGS code for item response theory. *Journal of Statistical Software, 36*(1), 1–34.

Dorans, N. J., & Kulick, E. M. (1986). Demonstrating the utility of the standardization approach to assessing differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*, 355–368.

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology, 72*, 19–29.

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.

Duanli, Y., von Davier, A. A. & Charles, L. (Eds.), *Computerized multistage testing: theory and application*. Boca Raton: Chapman and Hall/CRC.

Educational Testing Service. (2007). Innovations: news on research, products, and solutions for learning and education. Princeton, NJ: Author.

Embretson, S., & Reise, S. P. (2000). *Item response theory for psycho-logists*. New Jersey: Lawrence Erlbaum Associates.

Gierl, M. J., Lai, H., & Li, J. (2011). *Evaluating the performance of CATSIB in a multi-stage adaptive testing environment* (Final report submitted to Dr. Krista Breithaupt, Director of Research and Development of Medical Council of Canada). Vancouver, BC: Center for Research in Applied Measurement and Evaluation, University of Alberta.

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

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

Hick, W. E. (1951). Information theory and intelligence tests. *British Journal of Psychology, 4*, 157–64.

Holland, W. P., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.

Jiang, H., & Stout, W. (1998). Improved type I error control and reduced estimation bias for DIF detection using SIBTEST. *Journal of Educational and Behavioral Statistics, 23*, 291–322.

Kaira, L. T., & Sireci, S. G. (2010). Evaluating content validity in multistage adaptive testing. *CLEAR Exam Review, 21*(2), 15–23.

Linn, R., Rock, D., & Cleary, T. (1969). The development and evaluation of several programmed testing methods. *Educational and Psychological Measurement, 29*, 129–146.

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

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

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

Luecht, R. M. (2013). An introduction to assessment engineering for automatic item generation. In M. J. Gierl & T. M. Haladyna (Eds.), Automatic item generation: theory and practice (pp. 59–76). New Yourk: Routledge.

Luecht, R. M., & Nungester, R. J. (2000). Computer-adaptive sequential testing. In W. J. Vander Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: theory and practice* (pp. 117–128). Boston: Kluwer-Nijhof Publishing.

Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement, 35*, 239–249.

Makransky, G., & Glas, C. A. W. (2013). Modeling differential item functioning with group-specific item parameters: a computerized adaptive testing application. *Measurement, 46*, 3228–3237.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London: Champan and Hall/CRC Press.

Mead, A. (2006). An introduction to multistage testing [special issue]. *Applied Measurement in Education, 19*, 185–260.

Melican, G. J., Breithaupt, K., & Zhang, Y. (2010). Designing and implementing a multistage adaptive test: the Uniform CPA Exam. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 167–190). NW: Springer.

Nandakumar, R., & Roussos, L. (1997). *CATSIB: a modified SIBTEST procedure to detect differential item functioning in computerized adaptive testing*. Newton, PA: Law School Admissions Council.

Development Core Team, R. (2005). *R: a language and environment for statistical computing*. Vienna: The R Foundation for Statistical Computing.

Rotou, O., Patsula, L., Steffen, M., & Rizavi, S. (2007). *Comparison of multistage tests with computerized adaptive and paper-and-pencil tests [research report no. RR-07-04]*. Princeton, NJ: Educational Testing Service.

Roussos, L. (1996). *A type I error rate study of a modified SIBTEST DIF procedure with potential application to computerized-adaptive tests. Paper presented at the annual meeting of the Psychometric Society*. Canada: Alberta.

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

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.

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

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (1990). *Computerized adaptive testing: a primer*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Walker, C. M., Beretvas, S. N., & Ackerman, T. A. (2001). An examination of conditioning variables used in computer adaptive testing for DIF analyses. *Applied Measurement in Education, 14*, 3–16.

Weiss, D. J., & Betz, N. E. (1973). *Ability measurement: conventional or adaptive? (Research report, 73–1)*. Minneapolis: University of Minnesota, Psychometrics Methods Program, Department of Psychology.

Wood, R. (1973). Response-contingent testing. *Review of Educational Research, 43*, 529–544.

Yan, D., von Davier, A. A., & Lewis, C. (2014). *Computerized multistage testing: theory and applications*. UK: Taylor & Francis Group.

Zenisky, A., Hambleton, R. K., & Luecht, R. M. (2010). Multistage testing: issues, designs, and research. In E. J. van der Linden & C. E. W. Glas (Eds.), *Elements of adaptive testing* (pp. 355–372). New York: Springer.

Zheng, Y., Nozawa, Y., Gao, X., & Chang, H.-H. (2012). Multistage adaptive testing for a large-scale classification test: the designs, automated heuristic assembly, and comparison with other testing modes (ACT research reports 2012–6). Retrieved from http://media.act.org/documents/ACT_RR2012-6.pdf. Accessed 15 Dec 2016.

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, Canada: 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.

Zwick, R., & Thayer, D. T. (2002). Application of an empirical Bayes enhancement of Mantel-Haenszel differential item functioning analysis to a computerized adaptive test. *Applied Psychological Measurement, 26*, 57–76.