Language Testing in Asia

RESEARCH Open Access

Assessing the validity of an IAU General English Achievement Test through hybridizing differential item functioning and differential distractor functioning

Mehri Jamalzadeh[1], Ahmad Reza Lotfi[1*] and Masoud Rostami[2]

* Correspondence: lotfi.ahmadreza@gmail.com
[1]Department of English Language, Isfahan (Khorasgan) Branch, Islamic Azad University, Isfahan, Iran
Full list of author information is available at the end of the article

Abstract

The current study sought to examine the validity of a General English Achievement Test (GEAT), administered to university students in the fall semester of 2018–2019 academic year, by hybridizing differential information (DIF) and differential distractor function (DDF) analytical models. Using a purposive sampling method, from the target population of undergraduate students studying in different disciplines at Islamic Azad University (IAU), (Isfahan branch), a sample of 835 students taking GEAT were selected. The 60-item multiple-choice test comprised four sub-sections; namely, vocabulary, grammar, cloze test, and reading comprehension. The students' test scores served as the targeted data and the validity of the test was examined through the application of Cochran-Mantel-Haenszel (CMH) and multinomial log-linear regression models for detecting DIF and DDF, respectively. To account for the assumption of uni-dimensionality, the test sub-sections were analyzed independently. Furthermore, the assumption of local independence was checked based on correlational analysis and no extreme values were observed. The results of the study identified five moderate-level DIF items and one DDF item signaling an adverse effect on test fairness due to the existing biased items. Notably, these findings may have important implications for both language policymakers and test developers.

Keywords: Biased items, Differential item functioning (DIF), Differential distractor functioning (DDF), General English Achievement Test (GEAT), Test fairness

Introduction

One of the most significant discussions in language testing, which is becoming increasingly difficult to ignore, has been the question of validity. In the past few decades, the conceptualization of validity has undergone drastic changes and has left its initial orientations behind by focusing mainly on the question of whether the interpretations and actions based on test scores are justified in terms of evidential and/or consequential bases underlying test use (Messick, 1989).

Bond (2003, p. 179) states that validity should be considered "the core of any form of assessment that is trustworthy and accurate.". Therefore, it has to be borne in mind that validity, as an evolving complex concept, is closely related to the inferences made from assessment results. This requires inference-based evaluative judgments that are reflective of truth and lead to specific interpretations and actions.

According to Messick (1989, p. 5), "…what is to be validated is not the test or observation device but the inferences derived from the test scores or other indicators…". Largely because of the importance of precise and accurate inferences, test developers' precision in constructing tests has a great bearing on the validity of assessments so that the appropriateness of the inferences made about the results of a test reflects the appropriacy of the conclusions derived from the testees' performance on the items comprising a specific test (Messick, 1989, p. 6).

As a consequence, test items are written to measure psychological attributes which are often not directly possible. In fact, they serve as proxy measures of an unobservable psychological trait, a specific kind of knowledge, or psychomotor skill. Notably, test items require examinees to employ their intellectual and thinking skills in order to answer the test items. This provides test developers with a tangible yardstick by which they can improve the validity of the test and the quality of the inferences they make in order to judge the examinees' behavior in terms of answering the test items or performing the required skills (McNamara & Roever, 2006).

Differently stated, test items act as stimuli whose main purpose is to prompt a prescribed or expected answer. The endorsement to a particular test item can be representative of the fact that the examinee has acquired the intended trait or the attribute, or has the ability to perform the skill taught. Since tests are mainly utilized for making high-stake decisions about the examinees, the assessment of test results must be under careful examination and must be as fair as possible (Fulcher & Davidson, 2007; Shohamy, 2001; Stobart, 2005; Weir, 2005).

Clearly, potentially biased test items may adversely affect test fairness and might have significant implications for policymakers, test developers, and test-takers. Therefore, in developing a high stake test, test developers should determine the extent to which a test item is affected by bias or impact.

The item bias and item impact are closely tied to item validity and play a pivotal role in language testing. Item bias refers to the misspecification of the latent ability space, where items measuring multiple abilities are scored as though they are measuring a single ability. More specifically, according to Ackerman (1992), when two groups taking an identical test possess different multidimensional ability distributions and the test items can potentially differentiate these levels of abilities on such multiple dimensions, then any unidimensional scoring method would inadvertently result in item bias. In essence, item bias is an artifact of the testing procedure and is created when the source of the differential functioning of the item is irrelevant to the purpose of the test or the interpretation of the measure just because the item is tapping a factor which is over and beyond the targeted factor.

On the other hand, item impact exists when one group of examinees tends to answer a particular test item more correctly than the other group of examinees because the two groups truly differ on the underlying ability. In other words, item impact occurs when the item measures a relevant characteristic of the test without considering the actual differences existing between the two groups under assessment (Gelin, Carleton, Smith, & Zumbo, 2004).

Clearly, the consequential matters of test fairness and equity are quintessentially important because all examinees should enjoy equal opportunities to perform satisfactorily on a large-scale assessment, and hence being treated equitably in terms of their test scores (Moghadam & Nasirzadeh, 2020). The distinction between item bias and item impact is defined and clarified by the purpose of the measure. Therefore, test developers should carefully analyze the test items to see that they are flagged as displaying Differential item functioning (DIF). It is interesting to note that DIF is not the direct indicator of bias in a test. Rather, As Karami (2011) maintains, DIF is evidence of bias if the factor creating it is not relevant to the construct characterizing the test. In short, if that factor is part of the construct, it is preferably called item impact instead of bias.

In view of the above remarks, most researchers have focused on DIF and differential distractor functioning (DDF) separately. However, the present study aims to critically examine the effect of hybridizing DIF and DDF to improve the validity of university language achievement tests. Therefore, this study sought to investigate the extent to which integrating DIF and DDF analyses can mitigate test bias and improve fairness in assessment tests. Doing so will firstly fill in the gap from which the literature of the related topic suffers. Moreover, the findings of this study will help test developers not only to become aware of some apparently invisible biases but to avoid them and subsequently to develop tests with much higher validity and greater potential for fairly testing the language skills of the examinees.

To answer this question, this paper aims to provide a comprehensive review of the literature concerning certain key issues related to DIF and DDF which have a great bearing on the validity of assessment tests. Subsequently, by describing the methodological design by which the target research question can be operationalized, it will move to the results section reporting the obtained data in terms of different research variables. The results section will then be followed by the discussion of the findings in light of the research efforts presented and reported by other researchers with the same common goals and areas of interest. Finally, in the conclusion section, the paper will be brought to its final touch down by summarizing the main issues, suggesting possible implications, limitations, and the need for further research.

Literature review

In the past few decades, DIF has turned into an increasingly important area in language testing research. The analysis has frequently been used in psychometric circles for pinpointing the sources of bias at the item level (Zumbo, 1999). Through the application of DIF analysis, concerned practitioners in language testing have investigated test fairness and equity to explain factors such as misinterpretation of test scores, sexist or racist content, unequal prediction of criterion performance, unfair content with respect to the experience of test-takers, inappropriate selection procedures, inadequate criterion measures, and threatening atmosphere and conditions of testing studies (Brown, 1999; Jalili, Barati, & Moein Zadeh, 2020; Karami, 2011; Kim, 2001; Pae, 2004; Takala & Kaftandjieva, 2000; Walker & Göçer Şahin, 2020).

According to McNamara and Roever (2006), DIF was originated in the early twentieth century and was eventually used for the role of fairness in different tests to measure DIF. It was mainly triggered by researchers' interest in tapping social equity (Angoff, 1993). The main purpose of DIF was to specify the confounding variables through purging items that tinted the examinees' performance on tests.

Mellenbergh (1989), defining item bias as conditional dependence, suggests that statistical tests and indices based on item response theory may be used for detecting biased items when item characteristic curves of the two groups being tested do not coincide. By relying on empirical or simulated data and combining information on the regression of item responses on latent trait or observed test score and information on the latent trait or observed test score distribution, it is possible to identify the biased items.

Messick's (1980, 1989) debate on the consequential aspects of the tests in his validation framework oriented testing research toward such conceptual variables as DIF, validity, and fairness triggering a number of techniques for detecting biased items in different tests. As such, detecting differentially functioning techniques turned into a primary concern in test development and test use whose main objective was to demonstrate that the interpretations and uses made of test scores are credible and trustworthy (Geranpayeh & Kunnan, 2007).

Differential performance of different test-taking groups has a great bearing on the test development and test use procedures. Consequently, test users are responsible for ensuring that their test is free of bias and it provides a fair assessment. It is clear that the socio-ethical consequences of test use are particularly serious for high stakes tests. In fact, test fairness and test bias have a symbiotic relationship because when a given test is biased, it lacks fairness. With the emergence of critical language testing (CLT), it was contended that all uses of language tests are politically motivated, because as Shohamy (2001) argues, tests are in essence designed to manipulate society by imposing the will of the system on individuals.

Apparently, when groupings are involved, the likelihood that certain test items favor one group rather than the other is considerably high. Under such circumstances, the test may lack fairness for the disfavored group (Karami, 2011).

One of the key tests frequently used in the academic circles is the General English Achievement Test that is developed to measure skills and knowledge learned in a given grade level, usually through planned instruction, such as training or classroom instruction. In other words, these tests serve as a kind of summative evaluation chiefly devised to measure how much of a language someone has learned with reference to a particular course of study or teaching program. In the Iranian language teaching context, these tests are generally used for university students with different majors.

Given the ubiquity of various groups taking the test from different disciplines, it is important for test developers to make sure that the interpretation made on the test scores are valid because if the sources of DIF are irrelevant to the construct being measured by the test, some test items may act as a source of bias and the validity of the test is under question. Therefore, differential item functioning (DIF) analysis can be used to detect item bias when examinees from different groups with equal ability do not have the same chance of answering an item correctly (Camilli, Shepard, & Shepard, 1994; Fulcher & Davidson, 2013).

A considerable amount of literature on DIF has been published during the past 30 years. In 1992, focusing on black and Hispanic examinees, Dorans and Holland analyzed the data from an edition of the SAT to illustrate how applying a standardization approach to comprehensive differential item functioning (Cdif). The findings revealed that the standardization approach could be used to uncover differential speededness of the targeted participants.

In another study, using item response theory (IRT) methods, Banks (2009) found out that adjustments could be made at the analytic level by freeing and fixing parameters based on findings of differential item functioning (DIF). The results further revealed that high-stakes testing may require item removal or separate calibrations to ensure accurate assessment.

In a different study, McKeown and Oliveri (2017) examined differential item functioning (DIF) of the Deep Approaches to Learning scale on the National Survey of Student Engagement (NSSE) for Asian international and Canadian domestic 1st-year university students. The results indicated that only 1 of the 12 items functioned differentially.

In another study which was also implemented in 2020, Walker and Göçer Şahin, using differential item functioning, tried to evaluate interrater reliability as an index to determine if two raters differ with respect to their rating on a polytomous rating scale or constructed response item. More specifically, they used differential item functioning (DIF) analyses to assess inter-rater reliability and compared it with traditional interrater reliability measures. The results indicated that DIF procedures appear to be a promising alternative to assess the interrater reliability of constructed response items, or other polytomous types of items, such as rating scales.

In a large scale study, Cascella, Giberti, and Bolondi (2020) used a quantitative approach based on differential item functioning analysis within the Item Response Theory framework to quantify gender differences in tackling specific mathematical test items. Focusing on a sample of 1400 items administered in INVALSI tests answered by 30,000 Italian students per year since 2008, they demonstrated that DIF analysis is an effective approach to exploring the gender gap in relation to specific constructs of math education.

Similarly, in 2020, Zhu and Aryadoust tried to evaluate fairness in the Pearson Test of English (PTE) Academic Reading test, which is a computerized reading assessment test based on differential item functioning (DIF) across Indo-European (IE) and Non-Indo-European (NIE) language families. Analyzing the data from 783 international test-takers who took the PTE Academic test, using the partial credit model in Rasch measurement and using two main types of DIF, they found that uniform DIF is created when an item consistently gives a particular group of test-takers an advantage across all levels of ability, and non-uniform DIF (NUDIF) exists when the performance of test-takers varies across the ability continuum. The results identified 3 NUDIF items out of 10 items across the language families and showed no significant mother tongue advantage. The post-hoc content analysis of items further suggested that the decrease of mother tongue advantage for IE groups in high-proficiency groups and lucky guesses of low-ability groups may have contributed to the emergence of NUDIF items.

It is interesting to note that differential item functioning (DIF) is generally integrated with and differential distractor functioning (DDF) approach in order to flag the biased test items. Deng (2020) claims that the relationship between DIF and DDF is causal rather than correlational. She adds that there is no evidential basis to prove whether DDF could exist without the occurrence of DIF. Thus, in most cases, the analysis of DDF follows the results provided by DIF.

Green, Crone, and Folk (1989) extended the terminology of DIF to differential distractor functioning (DDF) to refer to the group differences that demonstrate whether an item is functioning differently for the different subgroups. Various terminologies have been employed for the concept. As an illustration, Dorans, Schmitt, and Bleistein (1992) use the term comprehensive differential item functioning (CDIF) combining DIF analysis with DDF, or more recently Park and WU (2017) refer to it as differential options functioning (DOF). Like DIF, there are also two types of DDF; namely, uniform and non-uniform. In uniform DDF, there is a consistent DDF pattern across all distractors in all directions. Conversely, non-uniform DDF points to the direction that indicates the differential effects of a distractor for different ability groups (Tsaousis, Sideridis, & Al-Saawi, 2018).

The presence of a relatively constant DDF effect across all *distractors* provides evidence that the cause of the DIF is rooted in the properties of the correct responses. As a result, when DDF effects are constant, the analyst can target content review on the correct response options. Clearly, the occurrence of only one distractor displaying a substantial DDF effect provides evidence that the DIF effect is either initiated by the properties of a specific distractor or by the possible interaction between the distractor properties and the content of the item stem (Penfield, 2008).

Many statistical methods have been developed for detecting DIF. Mapuranga, Dorans, and Middleton (2008) have identified four main methods for detecting DIF: (1) expected item score methods, (2) nonparametric odds ratio methods, (3) generalized linear model relationship between DDF and DIF methods, and (4) IRT-based methods. Each of them makes different assumptions about null DIF and utilize various models for implementing DIF.

Other popular methods to detect DIF comprise the Mantel-Haenszel (MH) procedure (Holland & Thayer, 1988), the Logistic Regression procedure (Swaminathan & Rogers, 1990), the multiple indicators multiple causes (MIMIC) model (Jöreskog and Goldberger (1975), the item response theory likelihood-ratio test (IRT-LR) (Thissen, Steinberg, & Wainer, 1993), Lord's IRT-based Wald test (Lord, 1977, 1980), and a Randomization Test based on an R-square change statistic (Edgington & Onghena, 2007; Sen, 2014).

Apparently, investigating these methods is certainly beyond the scope of this paper. As a consequence, the Mantel Haenzel (MH) method was used in this research for two main reasons. First, it offers a well-researched summary statistic with optimality characteristics that fit well with the ethical consequences of tests. Second, it provides a cheap and easy way of computing DIF by meticulously focusing on the probabilities of a correct response in focal and reference groups with the same ability (Gómez-Benito, Hidalgo, & Guilera, 2010; Holland & Thayer, 1988).

In a recent study by Belzak and Bauer (2020), the writers proposed an alternative method for selecting anchors and identifying DIF. In this method, a machine learning technique called regularization was used to impose a penalty function during estimation to remove parameters that have little impact on the fit of the model. By comparing lasso regularization with the more commonly used likelihood ratio test method in a 2-group DIF analysis, they found out that when large amounts of DIF are present and sample sizes are large, lasso regularization enjoys better control of type I error than the likelihood ratio test method with little decrement in power.

Paulsen, Svetina, Feng, and Valdivia (2020) employed the cognitive diagnostic models (CDMs) to diagnose examinees' strengths and weaknesses in a variety of content areas. The findings demonstrated that the detection of DIF in CDMs reflects how different magnitudes and types of DIF interact with CDM item types and group distributions and sample sizes to influence attribute- and profile-level classification accuracy. The results further suggested that attribute-level classification accuracy is robust to DIF of large magnitudes in most conditions, while profile-level classification accuracy is negatively influenced by the inclusion of DIF. In fact, conditions of unequal group distributions and DIF located on simple structure items had the greatest effect in decreasing classification accuracy.

Chen, Liu, and Zumbo (2020) using a propensity score method for investigating differential item functioning, proposed a new alternative for detecting DIF on performance assessment data. The proposed DIF method involved two main stages. In the first stage, propensity score matching was used to eliminate preexisting group differences before the test, ideally creating equivalent groups as in a randomized experimental study. In the second stage, linear mixed effects models were being adopted to perform DIF analysis based on the matched data set. The findings revealed that the utility of DIF analysis in assessing the targeted high-stakes functional English language proficiency test. It was also concluded that the proposed method could be applied to different types of test performance assessments.

Similarly, DDF has also been assessed through the application of specific methods. In fact, log-linear model fitting (Green et al., 1989), standardized distractor (SD) analysis (Dorans et al., 1992), an item response theory (IRT) model-based approach using a likelihood ratio (LR) test (Thissen, Steinberg, & Gerrard, 1986), the 2PL-nested logit model (2PL-NLM) (Suh & Bolt, 2011), the nominal response model (NRM) (Penfield, 2008), and the non-linear regression model (Drabinova, Martinkova, & Zvara, 2018). Most of these methods were implemented to estimate both DIF and DDF. Although DDF analysis has played a secondary role in the study of test fairness, it fortifies and improves the implementation of DIF since the DDF patterns can yield valuable information about the causes of DIF (Kato, Moen, & Thurlow, 2007).

Using a comparative approach, Middleton and Laitusis (2007) explored the possible variations of distractor functioning between students with disabilities and those without disabilities. The results indicated that seven of the items assessing the participants' performance on the state's reading standards exhibited DDF, while only three of the items that assessed performance on the state's writing standards indicated DDF.

Like DIF studies, the investigation of DDF and its role in determining test bias has turned into one of the most significant current discussions in the past few decades and the issue has been investigated by a large number of concerned practitioners (Green, Crone, & Middlton, 1989; Jalili et al., 2020; Martinkova et al., 2017; Middleton & Laitusis, 2007; Tsaousis et al., 2018, among others). However, to the best of the authors' knowledge, very few studies have tried to examine the synergy between DIF and DDF to evaluate bias and fairness in assessment tests.

More specifically, so far very few studies have investigated the integration of DIF and DDF in analyzing general English achievement tests. Considering the large number of university students from different disciplines studying in various Iranian universities, the investigation and improvement of this test should receive top priority in language testing research. The test administered at IAU universities in the past two decades has a multiple choice format addressing learners' general knowledge of English vocabulary, grammar, and reading comprehension across different disciplines. Consequently, cleansing the existing bank of language achievement test items can indubitably be considered a rewarding research outcome.

Method

This section seeks to present a detailed account of the research methodology and its components by providing the necessary information about the participants, materials, and instruments as well as the procedures employed for data collection and data analysis.

Participants

The participants in the present study were 835 male and female sophomore students, studying in various disciplines at IAU(Isfahan branch) in the fall semester of 2018–2019 academic year. Their age range varied between 19 and 27 and spoke Persian as their first language. These students (63% female and 37% female), who enjoyed a similar sociocultural background, formed the target population of the study selected based on a purposive sampling method. Clearly, the sampling method also referred to as a judgmental or expert sample, is considered a nonprobability sampling whose main purpose is to produce a sample that can be logically assumed to be representative of the population.

Instruments

A general English Achievement Test (GEAT) is often a large-scale test that determines students' current level of language ability and possible need for further language instruction. As such, under the mandates of the Ministry of Science and Technology in Iran, all tertiary level educational programs have included an English course in their curriculum and instruction to help learners to acquire an acceptable level of general English.

However, the use of General English tests to measure performance levels and to use the results for making high-stakes decisions about learners requires objective planning, design, and validation procedures because such results can be used as part of a total portfolio of data guiding the ethical consequences of testing (Pan, 2015). It is important to note that the validity of the test was established based on expert opinion (Davies, Soon, Young, & Clausen-Yamaki, 2004).

The GEAT, administered in the fall semester of 2018–2019 academic year, was a 60-item multiple choice test in which the test questions were divided into four different but complementary parts: Vocabulary (25 questions), Grammar (15 questions), Cloze Test (10 questions), and Reading Comprehension (10 questions). Being used as a criterion-reference test offered as a final term examination, it is one of the obligatory modules used for all courses in the bachelor program for all majors.

For computing the reliability of the test, raw agreement and kappa statistics are rarely used in practical testing situations due to the fact that the examinees are required to take two tests. However, it is also possible to estimate reliability from a single test administration. jMetrik, the software used in the DIF analysis section of this research, computes Huynh's raw agreement and kappa statistics called "Decision Consistency." Table 1 indicates that the index for the total test was 0.92 presenting a high reliability.

As can be seen, Huynh's raw agreement index has estimated reliability values for all test parts. Clearly, the reliability value for the grammar section of the test is low (0.81), while it is quite high for vocabulary and reading sections (0.90).

Data collection procedures

The GEAT answer sheets, belonging to the targeted participants, were received from the examination office. The permission to get access to answer sheets was granted by the university authorities on condition that the confidentiality concerning students' personal data are in place.

Detecting DIF

In DIF analyses, first the groups need to be adjusted for overall performance with regard to the measured trait in order to prepare the ground for comparing their performance on the test items. In other words, in assessing the examinees' response patterns to specific test items, the comparison groups (e.g., males vs. females) are initially matched on the targeted construct being measured (e.g., general English achievement). In fact, the main objective of DIF analysis is to substantiate whether the items in a standardized test favor the reference group (e.g., males) or the focal group (e.g., females). The analysis may help researchers or test developers determine whether item responses are equally valid for the specified groups or not (Zumbo, 1999).

The software used for detecting DIF in the study was jMetrik, which is one of the open software programs capable of performing item response theory (IRT) analysis for two-category and multi-category items. This software application has a variety of tools for performing statistical and psychometric analyses (Meyer, 2014). In general, the software is capable of providing researchers with statistical and graphical procedures (e.g., frequencies, correlations, bar charts, histograms, and kernel densities). Consequently, the answers given to a 60-item general English achievement test by a total of 835 students from different disciplines were analyzed to see whether they show differentiating item functioning according to the gender.

jMetrik employs the Cochran-Mantel-Haenszel (CMH), which is a technique that produces an estimate of an association between an exposure and an outcome after adjusting for or taking into account confounding. The statistic is generally used for testing statistical significance utilizing different related pieces of information such as common-odds ratio, ETS delta statistic, and the standardized mean difference (SMD).

Clearly, the MH procedure assumes that the ratio of those answering a particular item correctly is equally divided between reference and focal groups across all ability levels. In other words, as Lai, Teresi, and Gershon (2005, p. 284) maintain, "instead of testing against a general alternative hypothesis of any difference in correct response rates between groups, this statistic tests against a particular alternative of a common

Table 1 Reliability estimates

Vocabulary	Grammar	Cloze test	Reading	Total
0.90	0.81	0.82	0.90	0.92

odds-ratio across all blocking, or matching, categories." In fact, one of the major functions of the MH analysis is to provide statistical tests of whether the odds ratios are homogeneous or heterogeneous across the prespecified strata based on estimating the odds ratio of the exposure variable adjusted for the strata variable.

Another equally important piece of information provided by the MH procedure is the ETS delta statistic, which is frequently used across a broad range of DIF topics, such as equating aggregate scores in language testing (Karkee & Choi, 2005; Michaelides, 2010). Interestingly, the description of ETS rules in terms of common odds ratio for identifying and classifying the magnitude of DIF items is necessary in order to classify items as "A," "B," and "C" categories. "A" indicates that DIF magnitude is high because it has a CMH p-value larger than 0.05, or a common odds ratio exactly between 0.65 and 1.53. "B" category refers to moderate DIF levels that do not belong to either category "A" or "C". Ultimately, "C" category represents minimum DIF levels where common odds ratio is less than 0.53 when the upper bound of the 95% confidence interval for the common odds ratio is less than 0.65, or larger than 1.89 when the lower bound of the 95% confidence interval is larger than 1.53 (Meyer, 2014).

Finally, the mean-difference approach was used to evaluate the conditional between-group difference in the expected value of the item response variable. In general, two statistics belong to this approach: The standardized mean difference (SMD) and the polytomous SIBTEST. In this study, the ETS classification scheme also employed SMD in order to account for the standard deviation of the item scores for the total group standard of deviation (SD). Therefore, to calculate SMD between two groups, the mean of one group is subtracted from the other and the result is divided by the standard deviation (SD) of the population from which the groups were sampled. Additionally, the SIBTEST was used for investigating the causes of differential item functioning (DIF) based on Rasch model. By systematically manipulating DIF levels across multiple versions of an item, the factors responsible for causing DIF are identified (Schmitt, Holland, & Dorans, 1993).

It is interesting to note that Rasch model, a psychometric statistic, was used to examine answers to the questions on GEAT in terms of the trade-off between (a) the respondent's abilities and (b) the item difficulty. Two important concepts in the application of Rasch measurement in language assessment research must be considered: unidimensionality and local independence. In a unidimensional data set, a single ability is taken into account for the differences in scores. Local independence, on the other hand, measures whether the responses given to each item on a test are independent of the responses given to all other items. The objective is to indicate that each item is assessed independently of all other items (Bond & Fox, 2015; DeMars, 2010; Fulcher & Davidson, 2013; Markus & Borsboom, 2013; Van der Linden, 2018).

An item is said to be unidimensional if the systematic differences within the item variance are only due to the latent ability being measured. This idea is used to test the unidimensionality of a set of items using the principle of local independence (Lazarsfeld, 1959a and b). According to this principle, a set of items is regarded as unidimensional if there are no correlated residuals between the items once the variance due to the latent construct is controlled for. Notably, for local independence, jMetrik includes an option to check the assumption of local independence with Yen's *Q*3 statistic, which is the correlation of residuals for a pair of items (Yen, 1984, as cited in

Meyer, 2014). For the assumption of local independence correlation analysis was utilized to check for extreme values.

Detecting DDF

Items showing DIF were analyzed to determine the ones exhibiting DDF. In this study, DDF analysis was accomplished through the difNLR package by R software (Team, 2013). This package can potentially detect differential distractor functioning (DDF) based on the multinomial log-linear regression model. Item characteristic curves (ICC) plots for the male and female groups' selection of the test options are provided in the results section of the paper.

Results

This section focuses on the main findings related to the targeted research question addressing the integration of DIF and DDF and its possible effect on improving test validity and enhancing test fairness.

DIF analysis results

In DIF analysis, it is desirable to identify and flag only items with salient DIF. The results of the current study provided by the DIF analysis of 835 participants studying different majors at IAU (Isfahan branch) demonstrated that out of sixty items only five (i.e., 8.33% of the items) could be regarded as items exhibiting DIF. Subsequently, these items were judged to be category B items which exhibited a moderate amount of DIF. More specifically, items 1 and 38 were identified as flagged B+ favoring the focal group—that is, female students (3.33%), while items 40, 45, and 58 were flagged B– favoring the participants in the reference group (i.e., male students by 0.05%). It is clearly observed from the data presented in Table 2 below that the item discrimination values for the items 1, 38, and 40 were 0.28, 0.32, and 0.44 respectively and they favored female students. By contrast, items 45 and 58 with item discrimination values equal to 0.23 and 0.45 were in favor of male test takers.

DDF results

The results related to DDF analysis revealed that only the distractors belonging to DIF item 58 indicated that the distractors in this item showed DDF toward either male or female students. The item characteristic curve (ICC) plot for the two groups' selecting their answers from among item 58 distractors were then examined.

Table 2 Items exhibiting DIF

Item No.	Subtest	Chi-square	*p* value	Class	Item difficulty (TCC)	Item difficulty (Rasch)	Item discrimination
1	V	6.15	0.01	B+	0.81	– 1.15	0.28
38	G	9.33	0.00	B+	0.52	0.50	0.32
40	G	6.61	0.01	B-	0.62	– 0.01	0.44
45	C	8.42	0.00	B–	0.37	1.28	0.23
58	R	9.75	0.00	B–	0.72	– 1.28	0.45

Regarding choice "C," the correct answer**,** the male examinees showed a higher probability for endorsing the key—that is, choice "C" across all ability levels. As the ability level increases, the difference between the two groups' probability of giving a correct answer becomes negligible. Concerning choice "A," the low ability female examinees, more than the male group, had a higher probability of endorsing this distractor. As ability level increased, both groups showed the same probability for endorsing distractor "A." In regard with choice "B," the results indicated that low ability male examinees had a higher chance of selecting this distractor. As their ability increased, both groups showed a lower probability for endorsing distractor "B." For choice "D" the female group, at low ability level, had a higher chance of choosing this distractor. In fact, as ability level increased, both groups showed a lower probability for endorsing distractor "D." Figure 1 depicts the ICC for the item under scrutiny.

Discussion

The main objective of administrating a general English achievement to undergraduate university students from different disciplines is to measure the basic skills and knowledge learned in a given field of study, usually through planned instruction. The scores obtained on this test are often used in IAU educational system to determine the level of students' knowledge of general English in predicting their

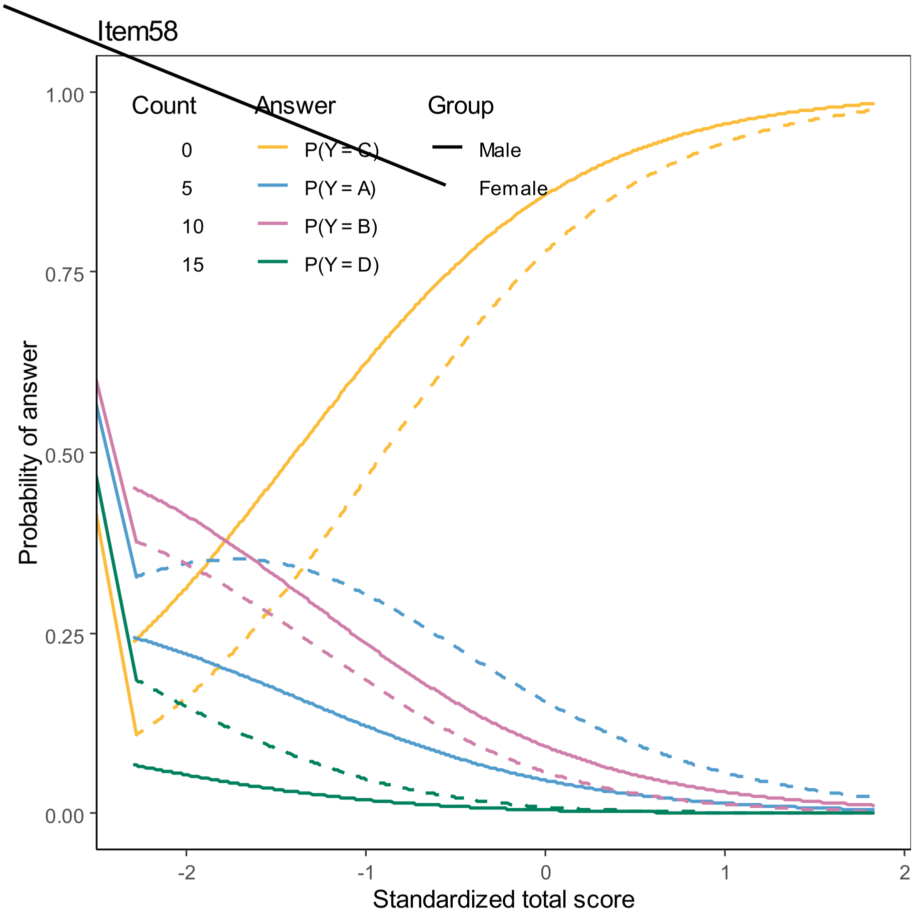

Fig. 1 Item characteristic curve plot for item 58

readiness in reading and comprehending English materials related to their fields of study. The application of appropriately constructed tests tapping the learners' general English knowledge is of paramount importance, and as a result, test developers must make sure that the information obtained from such examinations will be reliable and comparable.

This will only be achievable if the items used in the test do not function differentially among different sub-population of examinees across different disciplines because of factors which are not particularly relevant to the construct being measured. Under identical testing conditions, it is expected that the examinees from different groups with comparable ability levels exhibit similar probability of responding correctly to a given item. Under such circumstances, DIF represents a modern psychometric approach to the investigation of between-group score discrepancies. Conversely, DDF is used to investigate the quality of a measure through understanding biased responses across groups by shedding light on the potential sources of construct-irrelevant variance by examining whether the differential selection of incorrect distractors attracts various groups differently (Penfield, 2008, 2010).

Consequently, the present study set out with the aim of assessing the importance of hybridizing DIF and DDF analyses to detect the biased items in a 60-item, multiple choice General English Achievement Test (GEAT) administered at IAU, Isfahan branch in Iran. In fact, the principal objective was to integrate DIF with DDF in an attempt to potentially understand the causes of DIF and group biased responses on the targeted General English Achievement Test.

The results of this study revealed that integrating DIF analysis with the DDF approach to test assessment had a greater bearing on detecting the items that had an adverse influence on the validity of the test and its fairness. Largely because of the ethical consequences in high stakes tests like English achievement exams, the synergy between DIF and DDF is justified since such blending can help test developers to eliminate sources of uncertainty caused by inappropriately designed test items and distractors. On this basis, the present study sought to investigate to what extent hybridizing DIF analysis with the DDF approach could detect bias and improve the test fairness.

This study produced results which corroborate the findings of a great deal of the previous work in the field. For the sake of clarity, these works would be categorized as DIF studies, DDF studies, and DIF/DDF studies combined. Clearly, the results of the current study accord with earlier observations, which showed that the implementation of DIF analysis could practically improve the quality of test validation (Banks, 2009; Belzak & Bauer, 2020, Chen et al., 2020; McKeown & Oliveri, 2017; Paulsen et al., 2020; Zhu & Aryadoust, 2020, among many others).

In the same vein, the findings of the current study are consistent with the results reported by the authors focusing on the impact of DDF on test assessment. In fact, the findings support the efficacy of DDF approach to detecting whether the distractors on a multiple choice test functioned differently for the various groups of students and whether such test may be modified by removing inappropriate distractors while maintaining adequate test validity and information (Green, Crone & Folk, 1989; Jalili et al., 2020; Martinková & Drabinova, 2018; Middleton & Laitusis, 2007; Tsaousis et al., 2018, among others).

Finally, the findings of this study differ from Deng's (2020) report on the combined effects of DIF and DDF on test validation. In her research, she made an attempt to investigate the relationship between DDF and IDF as two critical ways for flagging test items with potential psychometric bias and detecting potential test fairness issues in multiple choice testing utilized for measuring the learners' achievements. Using multinomial logistic regression and binary logistic regression, she found that DIF and DDF are not correlated and there was no evidence to assume any association between the two.

A possible explanation for this might be that DIF occurs when examinees from different groups with different demographic background (e. g., gender or ethnicity) but the same underlying true ability have a different probability of answering the item correctly. On the other hand, differential distractor functioning (DDF) is a phenomenon when different distractors, or inappropriate option choices, attract various groups with the same ability differentially. Martinková and Drabinova (2018, p. 505) suggests that when "a given item functions differently for two groups, it is potentially unfair, thus detection of DIF and DDF should be routine analysis when developing and validating educational and psychological tests".

Hence, it could be hypothesized that the synergy between DIF and DDF has a great bearing on test validity. The DIF analysis of the 60-item IAUGEAT administered in the fall semester of 2018–2019 academic year revealed that five out of 6o items, namely items 1, 38, 40, 45, and 58, exhibited moderate DIF across gender. More specifically, while items 1 and 38 favored females, items 40, 45, and 58 were in favor of males. On the other hand, the results of DDF analysis indicated that 58 were sensitive to the analysis.

Therefore, it is interesting to note that the findings on the combined effects of DIF and DDF on test validation substantiate that these analytical approaches to test validation are interdependent. However, they serve as two quite independent procedures for tackling biased items so that test designers can replace biased items with more functionally logical ones. According to ETS (2016), to improve test fairness criteria, which are important guidelines for ameliorating the validity of tests, DIF, and DDF analyses, can help advance quality and equity in education. Thus providing fair and valid assessments has to be a top priority in high stakes testing all over the world.

The findings of the study have important implications for developing bias-free general English achievement tests. They are significant in at least one fundamental aspect: Validity is a multifaceted phenomenon. Analogously, validity can be considered a fortress which must be attacked from all sides and by all means. However, more research on this topic needs to be undertaken before the true nature of the synergy between DIF and DDF is more clearly understood.

Conclusion

The paper has given an account of and the reasons for the importance of test fairness by addressing the validity of a locally designed general English achievement test. In fact, this study set out to determine whether hybridizing DIF and DDF analyses of the students' responses to the items on the targeted general English achievement test could provide a better picture of test validity and fairness. The DIF analysis revealed that five items of the test exhibit moderate DIF and biased. In fact, while two items favored females, the other three were in the favor of male test-takers.

DDF analysis indicated that out of five items exhibiting moderate DIF, only one item, item 58 showed DDF. This item showed DDF in all three distractors. The combined effects of DIF and DDF analyses demonstrated that the test mostly favored the male test takers. Interestingly, in general, no specific hypothesis could be made about the behavior of these items.

The evidence from this study suggests that the social consequences of general English achievement tests are quintessentially important. Taken together, the results reflect that careful design and development of achievement tests is a prerequisite to test fairness which can be appreciably be enhanced by hybridizing DIF and DDF analytical methods. The current findings add substantially to understanding of test validation and the pivotal role it plays in the decisions made about the test-takers. More specifically, the study has gone some way toward enhancing test developers' awareness about the importance of hybridizing DIF with DDF. Clearly, DDF analysis is helpful in understanding the causes of DIF and provides information on whether DIF results from the test stem or the inherent features of the test distractors.

Finally, several limitations need to be considered in this study. The most important limitation lies in the fact that just one method was used to detect DIF and DDF even though there are numerous other methods by which these analytical indicators can be measured. Another limitation was related to the small number of items comprising the targeted test. With a small sample of items, caution must be applied, as findings might not be transferrable to other similar settings.

Therefore, if the debate is to be moved forward, a better understanding test validation and test fairness need to be developed. In fact, the future research should concentrate on investigating the synergy between DIF and DDF using several methods and larger samples of test items. Overall, the findings of this study may have a number of important implications for policymakers and test developers across the world.

Abbreviations

IAU: Isalamic Azad University; IAUGEAT: Islamic Azad University General English Achievement Test; DIF: Differential item functioning; DDF: Differential distractor functioning; CMH: Cochran-Mantel-Haenszel; MIMIC: The multiple indicators multiple causes; Cdif: Comprehensive differential item functioning; DOF: Differential options functioning; LR: Likelihood ratio; IRT: Item response theory; 2PL-NLM: 2PL-nested logit model; NRM: Nominal response model; ICC: Item characteristic curve

Authors' contributions

The first author was responsible for the design, data collection, and writing up of the paper. The second author was the supervisor of the first author's thesis out of which this article was extracted. He helped with clarifying ambiguities and reading and revising the article. The third author was the advisor. At the time of starting the study, he was the head of English department and responsible for general English. He helped with answering related questions. The authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

Data and material are available. They would be sent upon requesting through emailing.

Declaration

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of English Language, Isfahan (Khorasgan) Branch, Islamic Azad University, Isfahan, Iran. [2]English Department, Yazd University, Yazd, Iran.

Received: 10 September 2020 Accepted: 13 May 2021
Published online: 16 June 2021

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. https://doi.org/10.1111/j.1745-3984.1992.tb00368.x.

Angoff, W. H. (1993). *Perspectives on differential item functioning methodology*.

Banks, K. (2009). Using DDF in a post hoc analysis to understand sources of DIF. *Educational Assessment*, *14*(2), 103–118. https://doi.org/10.1080/10627190903035229.

Belzak, W., & Bauer, D. J. (2020). Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning. *Psychological Methods.*, *25*(6), 673–690. https://doi.org/10.1037/met0000253.

Bond, T. (2003). Validity and assessment: a Rasch measurement perspective. *Metodoliga de las Ciencias del Comportamento*, *5*(2), 179–194.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model*. https://doi.org/10.4324/9781315814698.

Brown, J. D. (1999). The relative importance of persons, items, subtests, and languages to TOEFL test variance. *Language Testing*, *16*(2), 217–238. https://doi.org/10.1177/026553229901600205.

Camilli, G., Shepard, L. A., & Shepard, L. (1994). *Methods for identifying biased test items*, (vol. 4). Sage.

Cascella, C., Giberti, C., & Bolondi, G. (2020). An analysis of differential item functioning on INVALSI tests, designed to explore gender gap in mathematical tasks. *Studies in Educational Evaluation*, *64*, 100819. https://doi.org/10.1016/j.stueduc.2019.100819.

Chen, M. Y., Liu, Y., & Zumbo, B. D. (2020). A propensity score method for investigating differential item functioning in performance assessment. *Educational and Psychological Measurement*, *80*(3), 476–498. https://doi.org/10.1177%2F0013164419878861. https://doi.org/10.1177/0013164419878861.

Davies, P. L., Soon, P. L., Young, M., & Clausen-Yamaki, A. (2004). Validity and reliability of the school function assessment in elementary school students with disabilities. *Physical & occupational therapy in pediatrics*, *24*(3), 23–43. https://doi.org/10.1300/J006v24n03_03.

DeMars, C. (2010). *Item response theory*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195377033.001.0001.

Deng, J. (2020). The relationship between differential distractor functioning (DDF) and differential item functioning (DIF): If DDF occurs, must DIF occur? (Doctoral dissertation, University of Kansas).

Dorans, N. J., Schmitt, A. P., & Bleistein, C. A. (1992). The standardization approach to assessing comprehensive differential item functioning. *Journal of Educational Measurement*, *29*(4), 309–319. https://doi.org/10.1111/j.1745-3984.1992.tb00379.x.

Drabinova, A., Martinkova, P., & Zvara, K. (2018). difNLR: DIF and DDF detection by non-linear regression models (Rpackage version 1.2.2). Retrieved from https://cran.r- project.org/web/packages/difNLR/index.html

Edgington, E., & Onghena, P. (2007). *Randomization tests*. CRC Press. https://doi.org/10.1201/9781420011814.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*. Routledge.

Fulcher, G., & Davidson, F. (2013). *The Routledge handbook of language testing*. Routledge. https://doi.org/10.4324/9780203181287.

Gelin, M. N., Carleton, B. C., Smith, M. A., & Zumbo, B. D. (2004). The dimensionality and gender differential item functioning of the mini asthma quality of life questionnaire (MINIAQLQ). *Social Indicators Research*, *68*(1), 91–105. https://doi.org/10.1023/B:SOCI.0000025580.54702.90.

Geranpayeh, A., & Kunnan, A. (2007). Differential item functioning in terms of age in the Certifi cate in Advanced English examination. *Language Assessment Quarterly*, *4*(2), 190–222. https://doi.org/10.1080/15434300701375758.

Gómez-Benito, J., Hidalgo, M. D., & Guilera, G. (2010). El sesgo de los instrumentos de medición. Tests justos. *Papeles del Psicólogo*, *31*(1), 75–84.

Green, B. F., Crone, C. R., & Folk, V. G. (1989). A method for studying differential distractor functioning. *Journal of Educational Measurement*, *26*(2), 147–160. https://doi.org/10.1111/j.1745-3984.1989.tb00325.x.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), Test validity (p. 129–145). Lawrence Erlbaum Associates, Inc.

Jalili, T., Barati, H., & Moein Zadeh, A. (2020). Using multiple-variable matching to identify EFL ecological sources of differential item functioning. *Journal of Teaching Language Skills*, *38*(4), 1–42.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *70*(351a), 631–639.

Karami, H. (2011). Detecting gender bias in a language proficiency test. *International Journal of Language Studies*, *5*(2), 27–38.

Karkee, T., & Choi, S. (2005). Impact of Eliminating Anchor Items Flagged from Statistical Criteria on Test Score Classifications in Common Item Equating. Online Submission.

Kato, K., Moen, R., & Thurlow, M. (2007). *Examining DIF, DDF, and omit rate by discrete disability categories*. University of Minnesota, Partnership for Accessible Reading Assessment. https://doi.org/10.1111/j.1745-3992.2009.00145.x.

Kim, M. (2001). Detecting DIF across the different language groups in a speaking test. *Language Testing*, *18*(1), 89–114. https://doi.org/10.1177/026553220101800104.

Lai, J. S., Teresi, J., & Gershon, R. (2005). Procedures for the analysis of differential item functioning (DIF) for small sample sizes. *Evaluation & The Health Professions, 28*(3), 283–294.

Lazarsfeld, P. F. (1959a). Latent structure analysis. In S. Koch (Ed.), Psychology: A study of a science (Vol. 3, pp. 476– 543). New York, NY: McGraw-Hill.

Lazarsfeld, P. F. (1959b). Problems in methodology.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge.

Lord, R. G. (1977). Functional leadership behavior: Measurement and relation to social power and leadership perceptions. *Administrative science quarterly*, *22*(1), 114–133. https://doi.org/10.2307/2391749.

Mapuranga, R., Dorans, N. J., & Middleton, K. (2008). A review of recent developments in differential item functioning. *ETS Research Report Series*, *2008*(2), i–32.

Martinková, P., Drabinová, A., Liaw, Y. L., Sanders, E. A., McFarland, J. L., & Price, R. M. (2017). Checking equity: Why differential item functioning analysis should be a routine part of developing conceptual assessments. *CBE—Life Sciences Education, 16*(2), rm2.

Martinková, P., & Drabinová, A. (2018). ShinyItemAnalysis for Teaching Psychometrics and to Enforce Routine Analysis of Educational Tests. *R Journal, 10*(2).

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. Routledge. https://doi.org/10.4324/9780203501207.

McKeown, S. B., & Oliveri, M. E. (2017). *Exploratory analysis of differential item functioning and its possible sources in the National Survey of Student Engagement*.

McNamara, T., & Roever, C. (2006). Psychometric Approaches to Fairness: Bias and DIF. *Language Learning, 56*(Suppl 2), 81–128. https://doi.org/10.1111/j.1467-9922.2006.00381.x.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International journal of educational research*, *13*(2), 127–143. https://doi.org/10.1016/0883-0355(89)90002-5.

Messick, S. (1980). Test validity and the ethics of assessment. *American psychologist*, *35*(11), 1012–1027. https://doi.org/10.1037/0003-066X.35.11.1012.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational researcher*, *18*(2), 5–11. https://doi.org/10.3102/0013189X018002005.

Meyer, J. P. (2014). *Applied measurement with jMetrik*. Routledge. https://doi.org/10.4324/9780203115190.

Middleton, K., & Laitusis, C. C. (2007). Examining test items for differential distractor functioning among students with learning disabilities. *ETS Research Report Series*, *2007*(2), i–34.

Michaelides, M. P. (2010) Sensitivity of Equated Aggregate Scores to the Treatment of Misbehaving Common Items. *Applied Psychological Measurement 34*(5):365-369

Moghadam, M., & Nasirzadeh, F. (2020). The application of Kunnan's test fairness framework (TFF) on a reading comprehension test. *Language Testing in Asia*, *10*(1), 1–21.

Pae, T. I. (2004). DIF for examinees with different academic backgrounds. *Language testing*, *21*(1), 53–73. https://doi.org/10.1191/0265532204lt274oa.

Pan, Y. C. (2015). Test impact: English certification exit requirements in Taiwan. *TEFLIN Journal*, *20*(2), 119–139. https://doi.org/10.15639/teflinjournal.v20i2/119-139.

Park, M., & Wu, A. D. (2017). Investigating differential options functioning using multinomial logistic regression. *International Journal of Quantitative Research in Education*, *4*(1-2), 94–119. https://doi.org/10.1504/IJQRE.2017.086511.

Paulsen, J., Svetina, D., Feng, Y., & Valdivia, M. (2020). Examining the impact of differential item functioning on classification accuracy in cognitive diagnostic models. *Applied Psychological Measurement*, *44*(4), 267–281. https://doi.org/10.1177/0146621619858675.

Penfield, R. D. (2008). An odds ratio approach for assessing differential distractor functioning effects under the nominal response model. *Journal of Educational Measurement*, *45*(3), 247–269. https://doi.org/10.1111/j.1745-3984.2008.00063.x.

Penfield, R. D. (2010). Modeling DIF effects using distractor-level invariance effects: Implications for understanding the causes of DIF. *Applied Psychological Measurement*, *34*, 1511–1565.

Schmitt, A. P., Holland, P. W., & Dorans, N. J. (1993). Evaluating hypotheses about differential item functioning. In P. Holland & H. Wainer (Eds.), Differential Item Functioning (pp. 281-316). Hillsdale, NJ: Lawrence Erlbauna.

Sen, S.(2014). Permutationtests. Retrieved from www.biostat.ucsf.edu/sen/statgen14/permutationtests.html.

Shohamy, E. (2001). Democratic assessment as an alternative. *Language testing*, *18*(4), 373–391. https://doi.org/10.1177/026553220101800404.

Stobart, G. (2005). Fairness in multicultural assessment systems. *Assessment in Education: Principles, Policy & Practice*, *12*(3), 275–287.

Suh, Y., & Bolt, D. M. (2011). A nested logit approach for investigating distractors as causes of differential item functioning. *Journal of Educational Measurement*, *48*(2), 188–205. https://doi.org/10.1111/j.1745-3984.2011.00139.x.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational measurement*, *27*(4), 361–370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x.

Takala, S., & Kaftandjieva, F. (2000). Test fairness: A DIF analysis of an L2 vocabulary test. *Language Testing*, *17*(3), 323–340. https://doi.org/10.1177/026553220001700303.

Team, R. C. (2013). R: A language and environment for statistical computing. https://www.R-project.org/

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin*, *99*(1), 118–128. https://doi.org/10.1037/0033-2909.99.1.118.

Thissen, D., Steinberg, L., & Wainer, H. (1993). *Detection of differential item functioning using the parameters of item response models*.

Tsaousis, I., Sideridis, G., & Al-Saawi, F. (2018). Differential distractor functioning as a method for explaining DIF: The case of a national admissions test in Saudi Arabia. *International Journal of Testing*, *18*(1), 1–26. https://doi.org/10.1080/15305058.2017.1345914.

Van der Linden, W. J. (2018). *Handbook of item response theory, three volume set*. CRC Press.

Walker, C. M., & Göçer Şahin, S. (2020). Using Differential Item Functioning to Test for Interrater Reliability in Constructed Response Items. *Educational and Psychological Measurement, 80*(4), 808–820.

Weir, C. J. (2005). *Language testing and validation*. Palgrave McMillan. https://doi.org/10.1057/9780230514577.

Zhu, X., & Aryadoust, V. (2020). An investigation of mother tongue differential item functioning in a high-stakes computerized academic reading test. *Computer Assisted Language Learning*, 1–25. https://doi.org/10.1080/09588221.2019.1704788.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF)*, (pp. 1–57). National Defense Headquarters.

Publisher's Note

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