International Journal of Assessment Tools in Education
2021, Vol. 8, No. 2, 376–393
https://doi.org/10.21449/ijate.730141
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Detecting Differential Item Functioning: Item Response Theory Methods Versus the Mantel-Haenszel Procedure

Emily Diaz [1,*], **Gordon Brooks** [2], **George Johanson** [3]

[1]Westat, Senior Research Associate, Education Studies
[2]Ohio University, Department of Educational Studies
[3]Ohio University, Department of Educational Studies

ARTICLE HISTORY
Received: Apr. 30, 2020
Revised: Mar. 11, 2021
Accepted: Apr. 04, 2021

Keywords:
Differential item functioning,
Monte Carlo simulation,
Type I error rate,
Mantel-Haenszel

Abstract: This Monte Carlo study assessed Type I error in differential item functioning analyses using Lord's chi-square (LC), Likelihood Ratio Test (LRT), and Mantel-Haenszel (MH) procedure. Two research interests were investigated: item response theory (IRT) model specification in LC and the LRT and continuity correction in the MH procedure. This study enhances the literature by investigating LC and the LRT using correct and incorrect model-data fit and comparing those results to the MH procedure. There were three fixed factors (number of test items, IRT parameter estimation method, and item parameter equating) and four varied factors (IRT model used to generate data and fit the data, sample size, and impact). The findings suggested the MH procedure without the continuity correction is best based on Type I error rate.

1. INTRODUCTION

In the field of psychometrics, item bias and test fairness are important issues that must be addressed (Kane, 2013). Item bias, differential item functioning (DIF), and impact are related but not synonymous (Zumbo, 1999). Item impact occurs when groups simply differ in performance on an item; when impact persists after controlling for overall skill on the construct being measured DIF is present; bias is pernicious DIF. Thus, DIF is the key to identifying possibly biased items.

Statistical tests of DIF are prone to both false positives (Type I errors) and false negatives (Type II errors. Roussos and Stout (1996) presented three reasons to research Type I error rates of DIF methods. First, removing a non-DIF item, or making a Type I error, unnecessarily wastes resources. Second, false positives explain why some testing organizations can neither understand nor ascertain the source of DIF in certain items. Finally, highly discriminating items can be mistakenly flagged for DIF (Li et al., 2012). Items with high discrimination indices contain higher information indices and are better able to discern differences between examinees

*CONTACT: Emily DIAZ emilydiaz@westat.com Senior Research Associate, Education Studies

ISSN-e: 2148-7456 /

with higher and lower levels of the underlying latent trait. Hence, false positives for these items are especially problematic and should not be needlessly removed.

1.1. Description of DIF Methods

According to Camilli and Shepard (1994) there are three theoretical reasons to prefer item response theory (IRT) methods over classical test theory (CTT) methods for DIF detection: item parameter estimates derived from IRT are less confounded and influenced with sample specific characteristics; IRT provides more accurate statistical properties of items than CTT to ascertain where the item functions differently (i.e., difficulty, discrimination, or pseudo-guessing); finally, the item characteristic curve (ICC) for each group can be graphed enhancing the understanding of items displaying DIF. According to Thissen et al. (1983) another advantage of IRT over CTT is that the fit between the data and the IRT model can be assessed statistically.

Lord's chi-square (LC) compares the performance of two groups on an item by examining item parameter differences depending on the specified IRT model (Lord, 1980). The group that is hypothesized to be favored, or have a higher probability of getting the item right, is the reference group (Camilli & Shepard, 1994; de Ayala, 2009). The group that is hypothesized to be disadvantaged, or have a lower probability of getting the item right, is the focal group (Camilli & Shepard, 1994; de Ayala, 2009). For LC, the item parameters are estimated separately for each group and are not directly comparable. Therefore, they need to be equated before meaningful comparisons can be made (Rupp & Zumbo, 2006; Stocking & Lord, 1983). LC follows a χ^2 distribution with degrees of freedom equal to the number of estimated parameters based on the IRT model implemented. Theoretically, LC is analogous to testing the equality of ICCs between the reference and focal groups. When the probability difference of getting an item right between the reference and focal groups is systematically the same across all ability levels, the item displays uniform DIF. Graphically, item characteristic curves for the groups are parallel (Camilli & Shepard, 1994). Non-uniform DIF occurs when the item favors one group over another for certain ability levels but reverses for other ability levels. Graphically, the item characteristic curves are not parallel. A benefit of using LC is that it can detect both uniform and non-uniform DIF.

The likelihood ratio test (LRT) assesses whether allowing the parameters for the studied item to vary across groups significantly improves the fit of the model. If so, then the studied item displays DIF. Judgments concerning fit are based on a comparison of the compact and augmented models. In the augmented model, an IRT model is fit such that all the item parameters are the same for the two groups except for the one item being studied, which varies across groups. In the compact model, the same IRT model specified in the augmented model is fit to the data such that all item parameters including the studied item are constrained to be the same in both groups (Thissen et al., 1988). The LRT test statistic is computed by $G^2 = -2LL_c - (-2LL_A)$ where $-2LL_c$ and $-2LL_A$ denote the negative two log-likelihood $(-2LL)$ of the compact and augmented models, respectively. The test statistic is compared to a χ^2 distribution with degrees of freedom equal to the number of estimated item parameters. An advantage of the LRT over LC is that item parameters are estimated together for both groups and do not need to be equated. However, a disadvantage is that the procedure takes a long time to implement because $n + 1$ models must be assessed for an n item test (Thissen et al., 1988). From a theoretical perspective, due to the asymptotic nature of the test statistic, the LRT and LC should yield the same conclusions provided the sample size is large ((Millsap & Everson, 1993; Thissen et al., 1993). This study adds to the literature by assessing this claim.

The Mantel-Haenszel (MH) procedure examines the relationship between item performance and group membership after taking into account total test performance (Dorans & Holland, 1993). This method examines whether item responses are independent of group membership

after controlling for observed score. The MH test statistic is compared to a χ^2 distribution with one degree of freedom and tests if the odds of members of the focal group getting the item right are the same as the odds of the reference group (Dorans & Holland, 1993). The MH statistic has been widely accepted because it is relatively easy to understand and implement, provides a χ^2 statistical significance test, and uses the odds-ratio as an effect size measure (Holland & Thayer, 1988; Millsap & Everson, 1993). Furthermore, an IRT model does not need to be fit to the data and the procedure does not require large sample sizes (Raju et al., 1993). One disadvantage of the MH is that it was designed to primarily detect uniform DIF (de Ayala, 2009; Millsap & Everson, 1993). However, in some cases MH can detect non-uniform DIF (Marañón et al., 1997; Mazor et al., 1994). Narayanan and Swaminathan (1996) note, that the MH procedure is ineffective in detecting non-uniform DIF that is also not ordinal.

1.2. Purposes of the Study

It is important to study DIF because certain measurement techniques require DIF analyses as a prerequisite (Shepard et al., 1985). For example, equating and test adaption are measurement approaches; that allow researchers to compare group estimates (i.e., item and/or person parameters) across separate test administrations, test forms, or groups (Cook & Eignor, 1991). When equating or adapting, truly biased items should not be present because these items are not measuring the concept similarly across groups. Hence, these items are uninformative and in fact can harm results(Kim & Cohen, 1992; Shepard et al., 1985).

Another import reason for studying DIF is that it addresses the validity of test score use because without it a test score is meaningless. In the United States the 1999 *Standards* (American Educational Research Association et al., 1999) called attention to test validity, which assesses whether a test is accurately measuring what it was designed to measure. According to the National Research Council (2007) in order to evaluate the trustworthiness and accuracy of score-based decisions testing companies must provide two types of evidence: the degree to which stated outcomes and purposes are achieved (i.e., intended effects) and the presence, or lack thereof, of adverse impact across groups of examinees. Furthermore, one particular type of evidence for validity is construct validity or the degree to which a test score is an accurate measure of the underlying latent variable it purports to measure (Creswell, 2009). According to Messick (1995) the value implications, interpretations, and meanings resulting from a test scores are a consequential aspect of construct validity. That is, when test scores are used in applied settings such as performance assessment, certification exam, licensure, course placement, college admittance, subject mastery and so forth there needs to be evidence of construct validity (Kane, 2009; Messick, 1995). In particular, DIF analyses statistically assess a potential threat to construct validity at the item level (Camilli, 2006; Kane, 2013).

When assessing DIF, there is a disparity between textbook presentations of IRT DIF methods and their frequency of use not only in practice but also in the Monte Carlo (MC) literature. IRT methods have a theoretical superiority to detect DIF (Camilli & Shepard, 1994; Thissen et al., 1993), yet they may not be as widely implemented in the simulation or MC literature on DIF as the MH and logistic regression procedures (Narayanan & Swaminathan, 1996). Raju (1990) commented that

> regardless of a particular investigator's decision for a given study, there is certainly a need for monte carlo [sic] and empirical studies to assess the degree of robustness and uniformity of item bias results obtained with the likelihood ratio, χ^2, and area procedures (p. 206).

This sentiment was again echoed by Raju et al. (1993) who stated that

> because this study was based on an empirical data set, it was not possible to know how many items were truly biased. There is obviously a need for a comprehensive Monte

> Carlo investigation to determine . . . the behavior of the IRT based methods with respects to false positives and false negatives (p. 310).

Despite these early calls, a recent (1/27/2021) search of the literature on *scholar.google.com* using the exact phrase terms "item response theory" and "differential item functioning" in the title returned 110 results, the majority of which focused on specific applications or software. When "Type-I" was added, there was only one citation, a dissertation concerning missing data. Another search using the terms "misspecification" and "item response theory" in the title returned only 2 results, neither related to DIF. The existing MC studies of DIF which have examined IRT and non-IRT DIF methods offer varied and sometimes conflicting research recommendations (Cohen et al., 1996; Cohen & Kim, 1993; DeMars, 2010; Kim et al., 1994; Herrera & Gómez, 2008; Lautenschlager & Park, 1988; Li et al., 2012; Lim & Drasgow, 1990; McLaughlin & Drasgow, 1987; Paek, 2010; Rudner et al., 1980; Sari & Huggins, 2015; Shepard et al., 1985; Wang & Yeh, 2003; Wells et al., 2009). Therefore, there still remain unknown aspects regarding these DIF methods such as IRT model fit, IRT model specification and misspecification, sample size, item discrimination variability, and item impact, which are addressed in this study and fill in the gap identified by Raju et al. (1993).

The main purpose of this study was to investigate and compare Type I error rates of DIF detection using LC, the LRT, and the MH procedures. Using multiple DIF methods, a form of psychometric triangulation, is a useful approach to investigate DIF in practice because each DIF detection method has different strengths and this adds to the research literature by allowing for comparisons to be made across DIF methods. Type I error was evaluated based on Bradley's (1978) stringent criterion interval [0.045, 0.055], which is equivalent to $\alpha \pm 0.1\alpha$, when $\alpha = .05$. Within the main purpose, two additional research interests guided this study: (1) the role of correct or incorrect IRT model specification in LC and the LRT, which was addressed using two simulations and (2) the role of the continuity correction in the MH procedure, which was addressed using one simulation. This MC study will add to and clarify the existing literature by determining the importance of correctly or incorrectly choosing the IRT model when computing LC and the LRT and comparing those results to the MH procedure. Correct and incorrect IRT model specification was added to enrich this study by providing guidance and recommendations to not only applied researchers but also to evaluators. In applied research determining the true and best IRT model to select when using LC and the LRT for a given dataset is never deterministically known (as it is in MC research) but is statistically assessed. Hence, these findings are useful to theoretical and applied researchers.

1.3. Variables in Monte Carlo DIF Studies

In the present study, the number of test items, IRT parameter estimation method, and item parameter equating were fixed while the IRT model used to generate data, IRT model used to fit the data, sample size, and impact varied based on the existing literature. For each DIF method, the MC literature surrounding the relationship between these variables of interest and the DIF method is discussed.

1.3.1. *Number of test items*

To obtain an accurate measure of ability for an individual, tests should include a sufficient number of items (Rogers & Swaminathan, 1993). From an IRT perspective, the minimum number of items needed to ensure accurate sampling error for item discrimination in the three-parameter logistic (3PL) model is 50 (Lord, 1968). From a CTT perspective, as test length increases standard error decreases resulting in a more accurate measurement of an examinees' ability using the observed score (Rogers & Swaminathan, 1993).

For much of the MC literature on DIF, the number of test items varied from 20 to 60. For LC, Wells et al. (2009) found test length did not influence Type I error rate while Cohen and Kim

(1993) found that Type I error was inflated for a 20 item test. For the LRT, Finch (2005) found an inconsistent relationship between Type I error and test length. For the MH procedure, Finch (2005) simulated test lengths of 20 and 50 items finding Type I error was consistently conservative but closer to the nominal level with the longer test. DeMars (2009) simulated three test lengths of 20, 40 and 60 items finding shorter tests led to higher Type I error rates for MH. Paek (2010) found Type I error for MH with the continuity correction was consistently conservative regardless of the number of test items. Similarly, Paek and Wilson (2011) found Type I error was consistently conservative regardless of the number of test items.

1.3.2. *Item parameter equating*

Item parameter equating is needed for LC only. The literature did not provide sufficient evidence to discern how item parameter equating influenced Type I error rate for two reasons. First, several studies fixed the item parameter equating technique (Kim & Cohen, 1995; Kim et al., 1994; Wells et al., 2009). Second, Candell and Drasgow (1988) investigate two equating methods but simulated impact in every cell of their design so their results could be confounded by impact. They used the weighted mean and sigma method developed by (Linn et al., 1981) and the test characteristic method of equating (Stocking & Lord, 1983) finding Type I error rate was consistently higher with this method.

1.3.3. *Sample size*

Sample size has varied widely in the MC DIF literature, ranging from 250 to 20,000. To obtain an accurate measure of the item parameters for both the 1PL and 2PL a sample size of 500 is adequate (Holland & Wainer, 1993; Sari & Huggins, 2015). For all three DIF methods the research literature suggests different and sometimes conflicting findings regarding the relationship between sample size and Type I error. For LC, there was no difference in Type I error across sample size (Wells et al., 2009). Kim et al. (1994) found more accurate results with larger sample sizes, while other studies found more accurate results with smaller sample sizes (Lim & Drasgow, 1990; McLaughlin & Drasgow, 1987). Finally, two studies did not find consistent results (Candell & Drasgow, 1988; Kim & Cohen, 1992).

For the LRT, Cohen et al. (1996) generated sample sizes of 250 and 1,000 while Stark et al. (2006) generated sample sizes of 500 and 1,000 both finding no marked differences in Type I error rates across sample sizes. Two additional studies found Type I error rate depended on whether group sample size was balanced or not. Finch (2005) found Type I error depended on the number of test items and group ability difference for the balanced condition. For the unbalanced condition Type I error was within the nominal level. Finch and French (2007) found with balanced group sample sizes of 250 and unbalanced group sample size results were within the nominal level. With balanced group sample size of 500, Type I error was conservative. However, with unbalanced sample size Type I error was inflated.

For the MH procedure, Narayanan and Swaminathan (1996) found Type I error was maintained across sample size. Other research has found conflicting results. Some studies have shown that larger sample sizes led to Type I error inflation (DeMars, 2009; Herrera & Gómez, 2008; Roussos & Stout, 1996) while other studies have shown both smaller and larger sample sizes resulted in conservative Type I error (Finch, 2005; Güler & Penfield, 2009; Herrera & Gómez, 2008; Paek 2010; Paek & Wilson, 2011; Rogers & Swaminathan, 1993). Due to the conflicting evidence and the need to obtain accurate item parameter estimates for the IRT DIF methods, the present study generated sample sizes of 500 and 1,000.

1.3.4. *Model misspecification*

The MC research literature was sparse concerning how IRT model selection affected Type I error rates. Most MC studies generated data and analyzed DIF based on fitting the true

underlying IRT model (e.g., Candell & Drasgow, 1988; Cohen & Kim, 1993; DeMars, 2009; Finch & French, 2007; Wang & Yeh, 2003). For LC, Lautenschlager and Park (1988) addressed model misspecification but the findings were not applicable to the current study because they generated multidimensional ability values (Lautenschlager & Park, 1988). For the LRT, Bolt (2002) addressed model misspecification for polytomous item response data finding that the IRT model selected impacted Type I error rate especially in the presence of impact.

The MH procedure uses the observed score to match the groups on ability so the underlying IRT model used to simulate data matters because the observed score is only a sufficient statistic for ability, θ, when the data follow the Rasch model and the one-parameter logistic (1PL) model (de Ayala, 2009; Zwick, 1990). Therefore, using the observed score in place of θ may cause problems for two-parameter logistic (2PL) and 3PL data. Narayanan and Swaminathan (1996) found Type I error was within Bradley's (1978) stringent criterion using 3PL data for all but one instance (i.e., reference group sample size of 500 and focal group sample size of 1,000). Roussos and Stout (1996) found Type I error was maintained when impact was not present but inflated when impact was present using 3PL data. Conversely, Rogers and Swaminathan (1993) found 2PL and 3PL data did not impact the number of Type I errors. The present study fit data using both the correct (same model used for data generation) and incorrect (different model used for data generation) IRT model. Note that when data created using the 1PL model are fitted using the 2PL model there is a case of overfitting and when data created using the 2PL model are fitted using the 1PL model there is a case of underfitting.

1.3.5. *Impact*

Impact occurs when the ability distribution of the groups being analyzed is not the same (Camilli, 2006; Camilli & Shepard, 1994; Clauser & Mazor, 1998; Dorans & Holland, 1993). Zumbo (1999) defined impact as different group probabilities of getting the item right because of true group ability differences on the underlying latent trait designed to be measured by the item. The MC research literature suggested an inconsistent relationship between impact and Type I error rate for LC, the LRT, and the MH procedure. In some instances, Type I error was conservative or maintained while in other cases it was inflated. For LC, Cohen and Kim (1993) did not find a clear relationship between impact and Type I error because their results depended on the nominal alpha level and estimation method. For the LRT, Finch and French (2007) found Type I error did not depend on impact. Finch (2005) found Type I error was generally closer to the nominal level when impact was present. Stark et al. (2006) found Type I error depended on impact and sample size. For MH, when simulating impact from 0.0 for 1.0 with intervals of 0.25 or 0.1 SD unit, Type I error increased as impact increased (DeMars, 2009; Li et al., 2012). However other studies found Type I error was conservative or maintained (Finch, 2005; Narayanan & Swaminathan, 1996; Paek, 2010; Paek & Wilson, 2011). The present study used three levels of impact: 0.0, 0.5, and 1.0.

2. METHOD

The open-source software R (R Core Team, 2013) was used to generate the data, run statistical analyses, and compute Type I error while BILOG-MG 3 (Zimowski et al., 2003) was used to estimate IRT models for LC and the LRT. In BILOG-MG, the number of cycles and quadrature points were both changed from the default of 10 to 20 and the number of Newton cycles was changed from the default of two to five to aid in more accurate item parameter estimates. The convergence criterion was changed from the default of 0.01 to 0.1 to aid model convergence for the LRT. Generally, the $-2LL$ value was greater than 1,000 so this small change did not greatly change the test statistic. In BILOG-MG, neither marginal maximum likelihood estimation (MMLE) nor Bayesian estimation can provide estimates for perfect items (proportion correct of 0.0 or 1.0). A condition was added to exclude datasets with perfect items.

Bayesian estimation, maximum marginal a posterior estimation, was chosen for parameter estimation. Four factors, sample size, group ability differences, IRT model used to generate data, and IRT model used to estimate item parameters, were manipulated in this study. The values selected for each factor were based upon theoretical and empirical rationale. This methodology fits the current trend for replication by providing sufficient detail, which will be discussed. It is at best difficult to compare the results of studies that do not provide sufficient, or sufficiently precise, details needed for replication or comparison. For all simulated conditions, the number of replications was fixed at 10,000 which is a relatively large number of replications as the number of replications in the literature ranges from one to 10,000 (Candell & Drasgow, 1988; Kim & Cohen, 1992; Li et. al, 2012).

In the present study, N_R and N_F denoted the number of examinees in the reference group and focal group, respectively. Two conditions, $N_R = N_F = 500$ and $N_R = N_F = 1{,}000$, were selected to represent moderate and large sample sizes. IRT DIF methods require larger sample sizes to accurately compute the variance-covariance matrix and the $-2LL$ value. Due to the complexity of computing these IRT DIF statistics, larger sample sizes were used.

Three levels of group mean ability difference, denoted μ_j, were manipulated. Theoretically, DIF analyses with group ability differences should not result in Type I error inflation, but prior research has shown that Type I error increased as impact increased (DeMars, 2009; Li et al., 2012). In this study, the reference group mean of the ability distribution was 0.0, 0.5, and 1.0 while the focal group mean of the ability distribution was fixed at 0. In all conditions SD was set at 1.0 for both groups.

A function was written in R to simulate dichotomous item response data (0 for incorrect and 1 for correct) based on a 50 item test with no DIF items for the reference and focal group separately with specified item parameters (a_i, b_i, and c_i) and person parameter (θ_j) following the 1PL and 2PL models. Test length was fixed at 50 items, the outer range of previous research (Cohen et al., 1996; Finch, 2005; McLaughlin & Drasgow, 1987). The higher number of items was simulated to obtain an accurate measure of ability for an individual and item difficulty and discrimination estimates for LC and the LRT (Lord, 1968; Rogers & Swaminathan, 1993). The 3PL model was not included as it would constitute another larger paper. The item difficulty parameter function inputs for the 1PL model were generated to follow a normal distribution while the pseudo guessing parameter was fixed at zero, which was consistent with prior research (Herrera & Gomez, 2008; Paek, 2010). For the 2PL model, the item discrimination parameter followed a normal distribution ($M = 1.1$, $SD = 0.25$), which was similar to Paek (2010) (i.e., a normal distribution with (M = 1.0, SD = 0.3)) This produced a range from 0.35 to 1.85 for 99% of values making it highly unlikely to encounter a negatively discriminating item. The item difficulty parameter followed a standard normal distribution, which was consistent with prior research (Herrera & Gomez, 2008; Paek, 2010). For the 1PL model, the item discrimination parameter was fixed at 1.1. This value was chosen for consistency because it was the mean of the item discrimination parameter in the 2PL model. Furthermore, this selection did not introduce any complications when comparing results across models. That is, if the item discrimination parameter had been chosen to be fixed at another value such as 0.8 or 1.2 it would have been more difficult to compare findings based on the IRT model due to the misalignment of item discrimination. In addition, this enhanced the generalizability of findings as data were generated from a different set of parameters each time as opposed to generating item response data based on a single test (Cohen et al., 1996; Sari & Huggins, 2015; Wang & Yeh, 2003). As previously noted, the number of items was fixed at 50 and no DIF items were simulated. For DIF detection, the choices for several parameters for data simulation are, admittedly, arbitrary. IRT model parameters were estimated using both the correct IRT model and incorrect IRT model.

Simulation I examined Type I error rates for LC and the LRT based on the 1PL and 2PL models under varied levels of sample size and impact when the fitting the correct IRT model. There were two types of correct model-data fit: (a) generating 1PL model data and fitting the 1PL model (hereafter denoted GEN1FIT1) and (b) generating 2PL model data and fitting the 2PL model (hereafter denoted GEN2FIT2). Fully crossing sample size, impact, and correct IRT model fit to data resulted in 12 cells for Simulation I, which are displayed in Table 1.

Table 1. *Summary of data collection procedure*.

Cell	IRT Model	Sample Size	Impact
1	1PL model	500	0.0
2	1PL model	500	0.5
3	1PL model	500	1.0
4	1PL model	1,000	0.0
5	1PL model	1,000	0.5
6	1PL model	1,000	1.0
7	2PL model	500	0.0
8	2PL model	500	0.5
9	2PL model	500	1.0
10	2PL model	1,000	0.0
11	2PL model	1,000	0.5
12	2PL model	1,000	1.0

To compute Type I error the first item was fixed and selected as the studied item reflecting previous studies (Güler & Penfield, 2009; Li et al., 2012; Roussos & Stout, 1996). Because the data are generated to be in random order, choosing to study the first item is equivalent to choosing a random item. The function *difLord* in the R package *difR* (Magis et al., 2010), was used for LC since simulation within R is advantageous for speed, efficiency, and potential replication. Item parameter estimates from BILOG-MG 3 were used as the inputs to compute LC for both the 1PL and 2PL models. A mean-sigma equating was used to place the focal group item parameter estimates onto the scale of the reference group (Cook & Eignor, 1991). Item parameter equating method was deliberately fixed to control both complexity of the study and the time needed to conduct the simulation. For the LRT, the $-2LL$ of the compact and augmented model denoted $L(C)$ and $L(A)$, respectively, from BILOG-MG were each saved as vectors in R. All the test items except the studied item were used as the anchor. The LRT was computed by comparing the difference of the two models ($G^2 = L(C) - L(A)$) to a χ^2 test with 1 *df* and 2 *df* for the 1PL and 2PL model, respectively. When -2*LL* differences were negative, implying the counterintuitive result that the compact model provided better fit, results were retained. The p value for these negative items was always 1.0 implying they were never rejections. This method was chosen for its consistency with results from IRTLRDIF (Thissen, 2001). As with LC, R was used to compute the LRT. For the MH procedure, the *difMH* function in R package difR (Magis et al., 2010) was used with the default of total score, or thin matching, to match the reference and focal group (item purification was judged unnecessary as no DIF items were simulated).

Simulation II examined Type I error rates for LC and the LRT based on the same levels of sample size and group ability difference used in Simulation I, but with the incorrect model fitted. Fully crossing sample size, impact, and incorrect IRT model fit to data also resulted in 12 cells for Simulation II. The only difference between Simulations I and II was whether the

specified IRT model was correctly fitted. There are two types of incorrect model-data fit: (a) generating 1PL model data and fitting the 2PL model (or overfitting, hereafter denoted GEN1FIT2) and (b) generating 2PL model data and fitting the 1PL model (or underfitting, hereafter denoted GEN2FIT1). Based on the previous literature, there was little guidance concerning how incorrect IRT model selection influenced IRT DIF analyses.

Lastly, Simulation III addressed the role of the continuity correction in the MH procedure. Simulation III examined Type I error rates in the MH procedure with and without the continuity correction under the same conditions, used in Simulations I and II. Given the conservative findings of Paek (2010), we included both forms of the MH procedure for completeness. Fully crossing sample size, impact, and IRT model used to generate data resulted in 12 cells for Simulation III. The code is available upon request from the authors.

3. RESULTS / FINDINGS

For Simulation I, the results for LC are displayed in Table 2. First, for GEN1FIT1, LC was consistently conservative, pretty stable, and not far from .05 regardless of sample size and impact using Bradley's (1978) stringent criterion. Second, for GEN2FIT2 Type I error rate increased as impact increased ranging from 0.042 to 0.098. Third, for GEN2FIT2 Type I error increased for LC as sample size increased regardless of impact. For GEN2FIT2 Type I error increased as both sample size and impact increased.

Table 2. *Type I error rates for LC and the LRT when fitting the correct IRT model.*

IRT Model	Sample Size	Impact	LC	LRT	MH	MH_CC
1PL model	500	0.0	0.041*	0.049	0.049	0.040*
		0.5	0.043*	0.056**	0.053	0.041*
		1.0	0.041*	0.064**	0.045	0.033*
	1,000	0.0	0.043*	0.050	0.050	0.042*
		0.5	0.040*	0.067**	0.048	0.040*
		1.0	0.043*	0.095**	0.050	0.042*
2PL model	500	0.0	0.042*	0.045	0.049	0.037*
		0.5	0.052	0.053	0.051	0.039*
		1.0	0.077**	0.059**	0.049	0.038*
	1,000	0.0	0.050	0.048	0.050	0.041*
		0.5	0.060**	0.063**	0.051	0.044*
		1.0	0.098**	0.075**	0.052	0.044*

Note. MH = the MH procedure without continuity correction; MH_CC = the MH procedure with continuity correction. Values marked with an * are conservative based on Bradley's (1978) stringent criterion; Values marked with ** are inflated based on Bradley's (1978) stringent criterion. The degrees of freedom are 1 and 2 for the 1PL and 2PL models, respectively.

The results for the LRT are also displayed in Table 2. First, when the groups were matched on ability Type I error was maintained, ranging from 0.045 to 0.050, for all four combinations of IRT model and sample size. Second, when the groups were not matched on ability (impact of 0.5 and 1.0) Type I error was inflated in all instances except one. The exception was GEN2FIT2 with a group sample size of 500 and impact of 0.5, which resulted in maintained Type I error. The actual Type I error rates ranged from 0.053 to 0.095 when impact was present. Third, Type

I error increased as sample size increased for all conditions in the LRT. The same results (conservative, maintained, or inflated) were seen across sample size in all but one condition. The exception was GEN2FIT2 with impact of 0.5. For this condition Type I error was maintained with a sample size of 500 but inflated with a sample size of 1,000.

For Simulation II, the results for LC are displayed in Table 3. First, in both cases of model misspecification Type I error for LC increased as impact increased across sample size ranging from 0.049 to 0.244. Second, in both cases of model misspecification Type I error rate for LC was consistently higher for the larger sample size condition compared to the smaller sample size condition. Third, Type I error was higher for GEN2FIT1 compared GEN1FIT2 in all combinations of sample size and impact except one. The exception was a sample size of 1,000 with impact of 0.0. Moreover, Type I error ranged from 0.041 to 0.091 compared to 0.044 to 0.244 based on GEN1FIT2 and GEN2FIT1, respectively. Fourth, the same Type I error results (conservative, maintained, and inflated) were seen for LC in all combinations of sample size and impact of GEN1FIT2 and GEN2FIT1 except one. The exception was a sample size of 500 and impact of 0.5 where Type I error was maintained for GEN1FIT2 but inflated for GEN2FIT1.

Table 3. *Type I error rates for LC and the LRT when fitting the incorrect IRT model.*

IRT Model Used to Generate Data	IRT Model Fit to Data	Sample Size	Impact	LC	LRT	MH	MH_CC
1PL model	2PL model	500	0.0	0.041*	0.044*	0.049	0.037*
			0.5	0.049	0.053	0.051	0.039*
			1.0	0.081**	0.055	0.049	0.038*
		1,000	0.0	0.047	0.048	0.050	0.041*
			0.5	0.059**	0.056**	0.051	0.044*
			1.0	0.091**	0.079**	0.052	0.044*
2PL model	1PL model	500	0.0	0.044*	0.073**	0.049	0.040 *
			0.5	0.081**	0.131**	0.053	0.041*
			1.0	0.152**	0.212**	0.045	0.033*
		1,000	0.0	0.045	0.100**	0.050	0.042*
			0.5	0.110**	0.227**	0.048	0.040*
			1.0	0.244**	0.347**	0.05	0.042*

Note. MH = the MH procedure without continuity correction; MH_CC = the MH procedure with continuity correction. Values marked with an *are conservative based on Bradley's (1978) stringent criterion; Values marked with ** are inflated based on Bradley's (1978) stringent criterion. The degrees of freedom are based on the IRT model fit to the data (i.e., 1 and 2 for the 1PL and 2PL models, respectively).

The results for the LRT are also displayed in Table 3 demonstrating one clear pattern: for all conditions Type I error rate increases as impact increases. There are five additional points worth noting. First, for GEN2FIT1 Type I error was inflated for all conditions of sample size and impact. Second, for GEN1FIT2 Type I error conclusions depended on sample size and impact. That is, with a sample size of 500 Type I error was conservative when groups were matched on ability but maintained when impact was present. However, with a sample size of 1,000 Type I error was maintained when the groups were matched on ability but inflated when impact was present. Third, Type I error rates were larger in GEN2FIT1 compared to GEN1FIT2 for all conditions of sample size and impact. Fourth, within each model misspecification category Type I error was higher for the larger sample size condition compared to the smaller sample

size condition. Fifth, the difference in Type I error rate from the larger sample size to the smaller sample size was larger in GEN2FIT1 compared to GEN1FIT2 regardless of impact.

For Simulation III, the results for the standard MH procedure using the continuity correction and the MH procedure without the continuity correction are given in Tables 1 and 2. In both tables the MH results are the same and were added to facilitate comparisons among the three DIF methods. For both forms of the MH procedure there was one consistent finding when using Bradley's (1978) stringent criterion across all simulated conditions: Type I error rates were conservative for the traditional MH procedure while Type I error rates were maintained for the MH procedure without the continuity correction. Furthermore, for both forms of the MH procedure impact, sample size, IRT model used to generate the data, and no combination of these three variables influenced Type I error rates to any great extent.

4. DISCUSSION and CONCLUSIONS

This study adds to the research literature by investigating IRT model specification or correct and incorrect model-data fit. Portions of the simulation results agree with previous research while other portions disagree. It is difficult to compare the results of this study with prior literature because many studies do not provide the methodological details needed for replication and comparison (e.g., two studies that examined LC [Lim & Drasgow, 1990; McLaughlin & Drasgow, 1987] did not mention item parameter equating). Overall, the results demonstrated two conclusions. First, when using large sample sizes of 500 and 1,000 per group regardless of impact and IRT model used to generate data the MH procedure is the preferred DIF method due to its Type I error performance consistency. Second, when using IRT DIF methods correct and incorrect IRT model specification and the effect of group differences cannot be ignored.

For LC in Simulation I GEN1FIT1 Type I error rates did not depend on sample size and impact using Bradley's (1978) stringent criterion, which is consistent with previous research (Wells et al., 2009). For GEN2FIT2, Type I error increased as sample size increased, which is consistent with research by Lim and Drasgow (1990) and McLaughlin and Drasgow (1987), but inconsistent with Kim et al. (1994). For GEN2FIT2 Type I error rate increased with impact, which did not agree with previous literature by Cohen and Kim (1993).

For the LRT in Simulation I Type I error was reasonably maintained when the groups were matched on ability for all four conditions of model-data fit and sample size. This was inconsistent with previous research (Finch, 2005; Finch & French, 2007; Stark et al., 2006). When the groups were not matched on ability (impact of 0.5 and 1.0) Type I error was inflated in seven of the eight model-data fit and sample size conditions. The exception was GEN2FIT2 with a group sample size of 500 and impact of 0.5 in which Type I error was maintained. These results were reasonably consistent with previous research (Finch, 2005; Stark et al., 2006). When impact was present with GEN2FIT2 Finch (2005) found Type I error was maintained with a group sample size of 500 while Stark et al. (2006) found Type I error somewhat conservative with a group sample size of 1,000. Although Type I error increased as sample size increased for all conditions, similar conclusions were generally made across sample size. This was inconsistent with some previous research (Cohen et al., 1996; Finch, 2005; Stark et al., 2006). There are several reasons why the present study may have inconsistencies with prior work. For example, the estimation methods differed for Cohen et al. (1996) and Finch and French (2007) and prior studies used 50-1,000 replications while the present study used 10,000.

For Simulation II, Type I error conclusions for GEN1FIT2 were generally consistent for LC and the LRT across the conditions with one exception. The exception was GEN1FIT2 with a group sample size of 500 and impact of 1.0 in which Type I error was inflated for LC and maintained for the LRT. However, for GEN2FIT1 the Type I error conclusions were only consistent when impact was present. Type I error increased as impact increased for all

conditions in both DIF methods. Type I error was generally inflated in both GEN1FIT2 and GEN2FIT1, but was more pronounced in GEN2FIT1. Finally, Type I error rates were lower for LC than the LRT in GEN2FIT1 but varied in GEN1FIT2. There was little research with which to compare these findings.

In Simulation III, Type I error rate was consistently somewhat conservative for the MH procedure. Impact did not lead to Type I error inflation when other variables were manipulated, which was both consistent and inconsistent with previous research (DeMars, 2009; Finch, 2005; Narayanan & Swaminathan, 1996; Paek, 2010; Roussos & Stout, 1996). Sample size did not influence Type I error rates when other variables were manipulated, which was also both consistent and inconsistent with prior research (DeMars, 2009; Herrera & Gómez, 2008; Narayanan & Swaminathan, 1996; Paek & Wilson, 2011; Roussos & Stout, 1996). The IRT model used to generate the data did not influence Type I error rates when other variables were manipulated. This was an interesting finding because the MH procedure matches the groups on observed score and not the underlying latent variable, θ. This finding was consistent with Rogers and Swaminathan (1993) who investigated how the 2PL model and 3PL model impacted the distributional shape of the MH test statistic and found no drastic differences in the number of Type I errors made for model-data fit. The finding also was consistent with Paek (2010) who simulated 1PL, 2PL, and 3PL data finding that Type I error was consistently conservative regardless of the IRT model used to generate data. This is noteworthy because the default method for the MH procedure in SPSS uses the continuity correction. Researchers and practitioners need to be mindful of this when interpreting their DIF results as they may be conservative. Furthermore, this study adds to the literature by extending the findings of Paek (2010). Paek (2010) examined the MH procedure under a variety of conditions while the present study included the MH procedure in conjunction with IRT DIF methods so that comparisons can be made between the two types of methods.

A key observation for the MH procedure was that no combination of IRT model used to generate the data, sample size, and impact, influenced Type I error rates to any great extent. This finding agrees with some previous research (Paek, 2010; Paek & Wilson, 2011), but was inconsistent with other research (DeMars, 2009; Herrera & Gómez, 2008; Narayanan & Swaminathan, 1996; Roussos & Stout, 1996).

4.1. Recommendations

There are six main recommendations based on this study. Recommendation one applies to statistical software for DIF analyses. Recommendations two through four apply to the results of the MH procedure and LC and the LRT using correct and incorrect model-data fit. The fifth recommendation compares the results of all three DIF procedures while recommendation six is a more general reflection on simulation studies.

First, it is important to empirically validate any R packages of interest prior to use. The authors of this study were not able to replicate the item parameter estimates from ltm at the time of data generation (Rizopoulos, 2006) which were used to implement LC. Thus, BILOG-MG was used in place of ltm.

Second, DeMars (2010) pointed out that when groups are not matched on ability Type I error can become inflated for the MH procedure. Previous research has shown this was true (DeMars, 2009; Li et al., 2012; Roussos & Stout, 1996). This study, however, demonstrates that this is not always the case. The Type I error for the MH procedure without the continuity correction was reasonably unaffected by group differences (impact) for all simulated conditions. This finding is important because the MH procedure is theoretically easy to understand, easy to implement, does not require knowledge of IRT, and is often used for DIF detection (Holland & Thayer, 1988; Wainer, 2010). Furthermore, this finding is particularly valuable to

psychometricians and applied researchers because it supports the use of the MH procedure for DIF analyses in the presence of impact based on Type I error rate.

Third, both studied IRT procedures generally showed inflated (and sometimes highly inflated) Type I errors with the combination of item impact and model misspecification. However, if the groups are matched on ability the LRT may be slightly preferred to LC when fitting the correct IRT model. Moreover, when the data are fit to the correct IRT model LC is often too conservative while the LRT is often too inflated based on Bradley's (1978) stringent criterion. When the data are fit to the incorrect IRT model, Type I error increased and became inflated as impact increased for both sample sizes. Therefore, choosing an appropriate IRT model for existing data is an important consideration (e.g., Bolt et al., 2014; Köse, 2014; Maydeu-Olivares, 2013) and, done well, can be arduous. In their chapter on the assessment of model-data fit, Hambleton et al. (1991) recommend a comprehensive set of procedures for assessing IRT model fit including checking model assumptions, parameter invariance, and model predictions. Furthermore, this IRT DIF finding is noteworthy because implementing LC based upon fitting the 1PL model is the only DIF procedure implemented in BILOG-MG 3. Therefore, psychometricians and applied researchers conducting DIF analyses using BILOG-MG 3 must be careful that their data correctly fit the 1PL model. That is, if a psychometrician or applied researcher is using BILOG-MG 3 for DIF analyses and the true underlying IRT model is the 2PL DIF results should be interpreted with caution as serious Type I error inflation can occur especially in the presence of impact.

Moreover, this recommendation is important because Type I error inflation is a serious problem as test items are expensive to construct. Luecht (2005) states that the: "ACPI [average-cost-per-item] typically runs from several hundred to more than fifteen hundred dollars per item" (p. 8). That is, making a Type I error by falsely removing a non DIF item is a serious financial consequence, which cannot be taken lightly.

Fourth, this study did not identify any unique advantage for using IRT methods over CTT methods based on Type I error rates. That is, based on Type I error rates the findings of this study do not support the theoretical advantages of using IRT for DIF analyses despite the recommendations of Camilli and Shepard (1994).

Fifth, taken together, recommendations two, three, and four agree with the principle of parsimony that the simpler method in comparison to the more complex method is better. That is, based on Type I error rate the MH procedure, a non-IRT based DIF method, should be use for DIF analyses instead of the more complex IRT DIF methods (LC and the LRT), which agrees with prior recommendations (Holland & Thayer, 1988; Wainer, 2010). Furthermore, this finding is particularly valuable to psychometricians and applied researchers because it supports a simpler method to implement and does not rely on correct IRT model selection. That is, the MH procedure overcomes the problem of Type I error inflation of LC and the LRT when selecting the incorrect IRT model.

Sixth, it is critically important that simulation studies in all areas provide sufficient detail for both comparison with prior research and replication. More bluntly, Monte Carlo research can be much more than a collection of case studies.

Acknowledgments

The authors received no financial support for the research, authorship, and/or publication of this article. Some pilot research that led to this manuscript was accepted and presented at the annual meeting of the American Educational Research Association in Philadelphia in April 2014.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research and publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

Authorship contribution statement

Emily Diaz: Investigation, Resources, Methodology, Visualization, Software, Formal Analysis, and Writing original draft. **Gordon Brooks**: Methodology, Supervision, Validation, and Writing orgnial draft. **George Johanson**: Methodology, Supervision, and Writing original draft.

ORCID

Emily Diaz https://orcid.org/0000-0001-9460-8647
Gordon Brooks https://orcid.org/0000-0002-2704-2505
George Johanson https://orcid.org/0000-0002-4253-1841

5. REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. American Educational Research Association.

Bolt, D. M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education, 15*(2), 113-141. https://doi.org/10.1207/S15324818AME1502_01

Bolt, D. M., Deng, S., & Lee, S. (2014). IRT model misspecification and measurement of growth in vertical scaling. *Journal of Educational Measurement, 51*(2), 141-162. https://doi.org/10.1111/jedm.12039

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*(2), 144-152. https://doi.org/10.1207/S15324818AME1502_01

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 220-256). American Council on Education.

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

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement, 12*(3), 253-260. https://doi.org/10.1177/014662168801200304

Cohen, A. S., & Kim, SH. (1993). A comparison of Lord's $\chi 2$ and Raju's area measures in detection of DIF. *Applied Psychological Measurement, 17*(1), 39-52. https://doi.org/10.1177/014662169301700109

Cohen, A. S., Kim, S. H., & Wollack, J. A. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement, 20*(1), 15-26. https://doi.org/10.1177/014662169602000102

Cook, L. L., & Eignor, D. R. (1991). IRT equating methods. *Educational Measurement: Issues and Practice, 10*(3), 37-45. https://doi.org/10.1111/j.1745-3992.1991.tb00207.x

de Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford Press.

Creswell, J. W. (2009). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. Sage.

DeMars, C. E. (2009). Modification of the Mantel-Haenszel and Logistic Regression DIF procedures to incorporate the SIBTEST regression correction. *Journal of Educational and Behavioral Statistics, 34*(2), 149-170. https://doi.org/10.3102/1076998607313923

DeMars, C. E. (2010). Type I Error inflation for detecting DIF in the presence of impact. *Educational and Psychological Measurement, 70*(6), 961-972. https://doi.org/10.1177/0013164410366691

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

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 29*(4), 278-295. https://doi.org/10.1177/0146621605275728

Finch, W. H., & French, B. F. (2007). Detection of crossing differential item functioning a comparison of four methods. *Educational and Psychological Measurement, 67*(4), 565-582. https://doi.org/10.1177/0013164406296975

Güler, N., & Penfield, R. D. (2009). A Comparison of the Logistic Regression and Contingency Table Methods for Simultaneous Detection of Uniform and Nonuniform DIF. *Journal of Educational Measurement, 46*(3), 314-329. https://doi.org/10.1111/j.17453984.2009.00083.x

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

Herrera, A. N., & Gómez, J. (2008). Influence of equal or unequal comparison group sample sizes on the detection of differential item functioning using the Mantel-Haenszel and logistic regression techniques. *Quality & Quantity, 42*(6), 739-755. https://doi.org/10.1007/s11135-006-9065-z

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

Kane, M. (2009). Validating the interpretations and uses of test scores. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 39-64). Information Age Publishing.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73. https://doi.org/10.1111/jedm.12000

Kim, S. H., & Cohen, A. S. (1992). Effects of linking methods on detection of DIF. *Journal of Educational Measurement, 29*(1), 51-66. https://doi.org/10.1111/j.17453984.1992.tb00367.x

Kim, S. H., & Cohen, A. S. (1995). A comparison of Lord's chi-square, Raju's area measures, and the likelihood ratio test on detection of differential item functioning. *Applied Measurement in Education, 8*(4), 291-312. https://doi.org/10.1207/s15324818ame0804_2

Kim, S. H., Cohen, A. S., & Kim, H. O. (1994). An investigation of Lord's procedure for the detection of differential item functioning. *Applied Psychological Measurement, 18*(3), 217-228. https://doi.org/10.1177/014662169401800303

Köse, I. A. (2014). Assessing model data fit of unidimensional item response theory models in simulated data. *Educational Research and Reviews, 9*(17), 642-649. https://doi.org/10.5897/ERR2014.1729

Lautenschlager, G. J., & Park, D. G. (1988). IRT item bias detection procedures: Issues of model misspecification, robustness, and parameter linking. *Applied Psychological Measurement, 12*(4), 365-376. https://doi.org/10.1177/014662168801200404

Li, Y., Brooks, G. P., & Johanson, G. A. (2012). Item discrimination and Type I error in the detection of differential item functioning. *Educational and Psychological Measurement, 72*(5), 847-861. https://doi.org/10.1177/0013164411432333

Lim, R. G., & Drasgow, F. (1990). Evaluation of two methods for estimating item response theory parameters when assessing differential item functioning. *Journal of Applied Psychology, 75*(2), 164-174. https://doi.org/10.1037/0021-9010.75.2.164

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement, 5*(2), 159-173. https://doi.org/10.1177/014662168100500202

Lord, F. M. (1968). An analysis of the verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement, 28*, 989-1020. https://doi.org/10.1177/001316446802800401

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

Luecht, R. M. (2005). Some useful cost-benefit criteria for evaluating computer-based test delivery models and systems. *Journal of Applied Testing Technology, 7*(2), 1-31.

Magis, D., Beland, S., Tuerlinckx, S., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42*, 847-862. https://doi.org/10.3758/BRM.42.3.847

Marañón, P. P., Garcia, M. I. B., & Costas, C. S. L. (1997). Identification of nonuniform differential item functioning: A comparison of Mantel-Haenszel and item response theory analysis procedures. *Educational and Psychological Measurement, 57*(4), 559-568. https://doi.org/10.1177/0013164497057004002

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives, 11*(3), 71-101. https://doi.org/10.1080/15366367.2013.831680

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1994). Identification of nonuniform differential item functioning using a variation of the Mantel-Haenszel procedure. *Educational and Psychological Measurement, 54*(2), 284-291. https://doi.org/10.1177/0013164494054002003

McLaughlin, M. E., & Drasgow, F. (1987). Lord's chi-square test of item bias with estimated and with known person parameters. *Applied Psychological Measurement, 11*(2), 161-173. https://doi.org/10.1177/014662168701100205

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749. https://doi.org/10.1037/0003-066X.50.9.741

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*(4), 297-334. https://doi.org/10.1177/014662169301700401

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement, 20*(3), 257-274. https://doi.org/10.1177/014662169602000306

National Research Council. (2007). *Lessons learned about testing: Ten years of work at the National Research Council*. The National Academies Press.

Paek, I. (2010). Conservativeness in rejection of the null hypothesis when using the continuity correction in the MH chi-square test in DIF applications. *Applied Psychological Measurement, 34*(7), 539-548. https://doi.org/10.1177/0146621610378288

Paek, I., & Wilson, M. (2011). Formulating the Rasch differential item functioning model under the marginal maximum likelihood estimation context and its comparison with Mantel-Haenszel procedure in short test and small sample conditions. *Educational and Psychological Measurement, 71*(6), 1023-1046. https://doi.org/10.1177/0013164411400734

R Core Team (2013). R: A language and environment for statistical computing. [Computer software]. R Foundation for Statistical Computing: Vienna, Austria. http://www.R-project.org/.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*(2), 197-207. https://doi.org/10.1177/014662169001400208

Raju, N. S., Drasgow, F., & Slinde, J. A. (1993). An empirical comparison of the area methods, Lord's chi-square test, and the Mantel-Haenszel technique for assessing differential item functioning. *Educational and Psychological Measurement, 53*(2), 301-314. https://doi.org/10.1177/0013164493053002001

Rizopoulos, D. (2006). Ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software, 17*(5), 1-25. https://doi.org/10.18637/jss.v017.i05

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*(2), 105-116. https://doi.org/10.1177/014662169301700201

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type I error performance. *Journal of Educational Measurement, 33*(2), 215-230. https://doi.org/10.1111/j.1745-3984.1996.tb00490.x

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980). A Monte Carlo comparison of seven biased item detection techniques. *Journal of Educational Measurement, 17*(1), 1-10. https://doi.org/10.1111/j.1745-3984.1980.tb00810.x

Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement, 66*(1), 63-84. https://doi.org/10.1177/0013164404273942

Sari, H. I., & Huggins, A. C. (2015). Differential item functioning detection across two methods of defining group comparisons: Pairwise and composite group comparisons. *Educational and Psychological Measurement, 75*(4), 648-676. https://doi.org/10.1177/0013164414549764

Shepard, L., Camilli, G., & Williams, D. (1985). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement, 22*(2), 77-105. https://doi.org/10.1111/j.1745-3984.1985.tb01050.x

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*(6), 1292-1306. https://doi.org/10.1037/00219010.91.6.1292

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201-210. https://doi.org/10.1177/014662168300700208

Thissen, D. (2001). *IRTLRDIF user's guide: software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning*. Retrieved from http://www.unc.edu/~dthissen/dl.html

Thissen, D., Steinberg, L., Pyszczynski, T., & Greenberg, J. (1983). An item response theory for personality and attitude scales item analysis using restricted factor analysis. *Applied Psychological Measurement, 7*(2), 211-226. https://doi.org/10.1177/014662168300700209

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147-169). Lawrence Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Lawrence Erlbaum.

Wainer, H. (2010). 14 conversations about three things. *Journal of Educational and Behavioral Statistics, 35*(1), 5-25. https://doi.org/10.3102/1076998609355124

Wang, WC., & Yeh, YL. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement, 27*(6), 479-498. https://doi.org/10.1177/0146621603259902

Wells, C. S., Cohen, A. S., & Patton, J. (2009). A range-null hypothesis approach for testing DIF under the Rasch model. *International Journal of Testing, 9*(4), 310-332. https://doi.org/10.1080/15305050903352073

Zimowski, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG 3: Item analysis and test scoring with binary logistic models*. [Computer software]. Scientific Software.

Zumbo, B. (1999). *A handbook on the theory and methods of differential item functioning: Logistic regression modeling as a unitary framework for binary and Likert-type item scores*. Directorate of Human Resource Research and Evaluation, National Defense Headquarters.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational and Behavioral Statistics, 15*(3), 185-197. https://doi.org/10.3102/10769986015003185