Research Report

A Review of Recent Developments in Differential Item Functioning

Raymond Mapuranga

Neil J. Dorans

Kyndra Middleton

August 2008

ETS RR-08-43

A Review of Recent Developments in Differential Item Functioning

Raymond Mapuranga, Neil J. Dorans, and Kyndra Middleton
ETS, Princeton, NJ

August 2008

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS' constituents and the field.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

In many practical settings, essentially the same differential item functioning (DIF) procedures have been in use since the late 1980s. Since then, examinee populations have become more heterogeneous, and tests have included more polytomously scored items. This paper summarizes and classifies new DIF methods and procedures that have appeared since the early 1990s and assesses their appropriateness for practical use. Widely used DIF methods are evaluated alongside these new methods for completeness, clarity, and comparability.

Key words: Differential item functioning, statistical criteria for differential item functioning, practical criteria for differential item functioning

Acknowledgments

The authors thank Michael Zieky, Anna Kubiak, Michael Jodoin, Elizabeth Stone, Adam Wyse, Daniel Eignor, Sandip Sinharay, and Adele Tan for useful advice and Kim Fryer for editorial help.

Table of Contents

1. Introduction

Essentially the same differential item functioning (DIF) procedures have been used in practice since the late 1980s. The use of new item formats, new item types, and new test administration procedures, as well as the increasing number of test takers with limited English language proficiency requires a re-examination of existing DIF procedures and a consideration of more recent developments. This paper summarizes and classifies new DIF methods and procedures while judging their appropriateness for practical use. The term *new* refers to methods that have appeared in the research literature since the 1990s, when several articles reviewing DIF methodology were authored (e.g., Millsap & Everson, 1993; Potenza & Dorans, 1995). While we have attempted to include all new methods, we may have missed some.

DIF analysis is an important step in evaluating tests for fairness and equity. DIF occurs when different groups of examinees with the same level of proficiency in a domain have different expected performance on an item. In DIF analysis, the sample is usually divided into two subgroups. The reference group typically provides a baseline for performance (e.g., White or male) and the focal group is typically the focus of fairness concerns (e.g., Black, Hispanic, or female).

The practice of analyzing DIF developed as a response to practices that confounded differences in item functioning with differences in score distributions (Zieky, 1993). Analysis of DIF is not a statistical or psychometric operation that is performed in a policy vacuum. Rather, DIF analysis is attuned to policy issues, and several focal groups are protected against unfair practices by legislation (e.g., female, Black, and Hispanic examinees and examinees with disabilities). Examples of legislation related to the need for DIF analyses are the Individuals with Disabilities Education Act (IDEA;, 1991, 1997), which relates to the testing of examinees with disabilities, and the Civil Rights Act of 1964, which covers the concept of fairness based on race, color, religion, sex, or national origin (Camilli, 2006). The next section reviews the challenges posed by the aforementioned testing contexts and populations.

2. Motivation for Review

Several enhancements to DIF analysis would be useful. First, it would be helpful to develop or identify methods that can efficiently detect DIF in tests containing constructed response (CR) items. Specifically, matching criteria for DIF analyses of CR items tend to be questionable and ineffectual. That is, the use of multiple choice scores as a matching criterion is

often inadequate, since the criterion sometimes has a low correlation with the CR score. Additionally, DIF analysis on CR items using a CR criterion score is problematic when the test includes only a few CR items, because of low reliability of measurement. Furthermore, there are no well-established procedures for correcting or improving CR questions that exhibit DIF. Because tests typically include only a few CR items, discarding these items could mean throwing away a significant portion of a test in situations where these items exhibit DIF.

Studies have also shown that contextual factors (e.g., language and cultural characteristics) can have an impact on DIF analyses, particularly for English language learners (ELL). For example, it was shown that when Hispanic examinees took the SAT® Reasoning Test (formerly the SAT Verbal test), items containing specific linguistic features (e.g., true cognates, false cognates, homographs) and items that are of special cultural interest exhibited DIF (Schmitt, Curley, Bleistein, & Dorans, 1988; Schmitt, Holland, & Dorans, 1993). DIF often occurs when some examinees use the language being tested as an *academic language* while others use the language as a *home language*. When native language speakers are tested in their native language, DIF occurs on some items between them and non-native speakers due to language familiarity learned outside the classroom.

Methods that are effective with small sample sizes would represent a substantial advancement in DIF analyses, because several situations exist where sample size is small. Specifically, for DIF analyses of some racial and ethnic subgroups, especially Native Americans, samples are typically too small (e.g., less than 200 per group as described in Clauser & Mazor, 1998). Another occurrence of small sample sizes used for DIF analyses occurs among examinees with disabilities (Stone, Cook, Cline, & Cahalan-Laitusis, 2007). If sample sizes are too small, the analysis may not have enough power to detect DIF (see Puhan, Moses, Yu, & Dorans, 2007 for a recent example).

Another concern pertaining to the inadequacy of DIF analyses has occurred in recent years because of rapidly changing U.S. demographics. Specifically, the increasing number of examinees who are not adequately proficient in English may mean that current approaches to DIF analysis are no longer robust. Because DIF samples are often selected based on those who indicate that English is their first language, it is important to evaluate how well these results generalize to the full test-taking population (Sinharay, Dorans, & Liang, 2008). Moreover, given other demographic changes as well, currently used matching criteria may not be adequate, and

technical improvements might be needed. In addition, another technical issue involving the matching criterion concerns defining circumstances in which matching variables should include the studied item.

The aforementioned challenges and areas of concern provided the motivation for our work on classifying recent DIF methods. Understanding of these challenges and concerns will help in identifying promising methods that can potentially address these issues. The remainder of the paper summarizes the most commonly used DIF methods and introduces technical, practical, and statistical classification criteria that will be used in their evaluation. Using the aforementioned criteria, the new methods are then classified. Lastly, the extents to which the new methods address the concerns that motivated this review are evaluated, and some suggestions and implications for future research are discussed.

3. Older Methods

A sizeable number of methods, analyses, and applications of DIF exist in the research literature. Currently, some of the most commonly used or studied older DIF procedures include the standardization (STAND)[1] procedure (Dorans & Kulick, 1986), SIBTEST (Shealy & Stout, 1993), the Mantel-Haenszel (MH) procedure (Holland & Thayer, 1988), and logistic regression (Swaminathan & Rogers, 1990). Because of their popularity, these are the only old methods that will be compared to the new ones described in this paper.

In prior, or older, research, different analysis procedures were proposed for reducing bias and estimation error in DIF analysis. These included iterative purification procedures under various conditions: (a) different matching criteria, (b) different analytic units (i.e., item-, bundle-, or testlet-level), and (c) different assumptions about tests' dimensional structures (e.g., Gierl, Bisanz, Bisanz, Boughton, & Khaliq, 2001; Mazor, Kanjee, & Clauser, 1995; Oshima, Raju & Flowers, 1997; Wainer, 1995). Furthermore, variations of the aforementioned procedures were developed to study non-uniform DIF (Li & Stout, 1996).

Additionally, the literature includes numerous comparison studies that used older DIF procedures, such as STAND, SIBTEST, MH, item response theory (IRT) based, and logistic regression, (Chang, Mazzeo, & Roussos, 1996; de Ayala, Kim, Stapleton, & Dayton, 2002; Zwick, Thayer, & Mazzeo, 1997). Moreover, some review papers have summarized different aspects and characteristics of several older DIF methods. For example, Clauser and Mazor (1998) summarized and evaluated technical aspects, while Haladyna and Downing (2004)

focused on the conceptual aspects of several DIF procedures. Another example is the Millsap and Everson (1993) paper, which reviewed bias detection methods used in educational and psychological measurement.

This paper builds on the Potenza and Dorans (1995) taxonomy, which summarized the most commonly used binary and polytomous DIF procedures and classified them according to two basic criteria: (a) associated matching variable (either an observed score or an unobserved latent trait), and (b) the assumed relationship between item scores and matching variables (i.e., parametric or nonparametric). Among binary DIF procedures, logistic regression was classified as both observed score and parametric; IRT-based procedures as both latent trait and parametric; MH and STAND methods as both observed score and nonparametric; and SIBTEST[2] as both latent trait and nonparametric. Among the polytomous methods, logistic regression was classified as both observed score and parametric; IRT-based methods as both latent trait and parametric; poly-SIBTEST (Chang et al., 1996) and generalized partial credit model DIF (Muraki, 1993) as both nonparametric [sic] and latent trait, and the STAND, generalized MH (Mantel & Haenszel, 1959; Zwick , Donoghue, & Grima, 1993), HW1 (Welch & Hoover, 1993), HW3 (Welch & Hoover, 1993), and Mantel (Dorans & Schmitt, 1993; Mantel, 1963; Zwick et al., 1993) procedures were classified as both observed score and nonparametric.

4. Classification Using DIF Criteria

In what follows, our DIF classification will include the following criteria: (a) null DIF, (b) studied item score, (c) matching variable, and (d) grouping variable. Each of these criteria will be explained below.

Null DIF is the absence of DIF. One definition of null DIF, observed-score null DIF, is that all individuals with the same score on a test should have the same proportions answering the item correctly regardless of whether they are from the reference or focal group. The latent-variable definition of null DIF compares the performance of focal and reference subgroups that are matched with respect to a latent variable.

The s*tudied item score* refers to the scoring rule used for the items being studied for DIF. Studied items can either be scored as correct/incorrect (i.e., binary) or scored using more than two response categories (i.e., polytomous).

The *matching variable* is a variable used in the process of comparing the reference and focal groups (e.g., total test score or subscore) so that comparable groups are formed. In other

words, matching is a way of establishing score equivalence between groups that are of interest in DIF analyses. The matching variable can either be an observed score or an unobserved latent variable and either a univariate or a multivariate variable.

In most DIF analyses, a single focal group is compared to a single reference group where the subgroup classification variable (i.e., gender, race, geographic location, etc.) is referred to as the *grouping variable*. This approach ignores potential interactions between types of subgroups, (e.g., male/female and ethnic/racial). Although it might be better to analyze all grouping variables for DIF simultaneously, (for statistical and computational efficiency) most DIF methods compare only two groups at a time (Zhang, Dorans, & Matthews-Lopez, 2005).

What follows is a classification and summary of DIF methods according to the following general groupings: (a) expected item score methods, (b) nonparametric odds ratio methods, (c) generalized linear model methods, and (d) IRT-based methods.

4.1 Expected Item Score Methods

The null-DIF definition of expected item score methods states that, at each level of the matching variable, there is no difference in proportions correct between the reference and focal groups. In other words, this definition implies that there is a zero difference in the expected item score given the matching variable (Potenza & Dorans, 1995). STAND was tabbed as the progenitor to SIBTEST by Shealy and Stout (1993), since both methods use the concept of expected item score. For this reason, the two methods will be evaluated alongside each other. STAND has undergone extensions that include polySTAND (Dorans & Schmitt, 1993) and Cdif (Dorans, Schmitt, & Bleistein, 1992) which were discussed by Potenza and Dorans (1995). This paper will revisit STAND and will also discuss two newer developments: smoothed STAND (Lyu, Dorans, & Ramsay, 1995) and DIF dissection (Zhang et al., 2005).

Over the years, several enhancements and new applications of SIBTEST have been documented. For example, SIBTEST was extended for use with item bundles (Douglas, Roussos, & Stout, 1996; Gierl et al., 2001; Gierl & Bolt, 2001) and for detection of DIF in polytomous items (Chang, Mazzeo, & Roussos, 1993, 1996). STAND and SIBTEST methods (i.e., SIBTEST, kernel smoothed SIBTEST and MULTISIB) are evaluated in Appendix A, and explanations of this evaluation are provided below.

STAND and smoothed STAND have the same DIF classifications because both these observed-score methods use binary scored items and a univariate matching variable. Both

methods also use a single grouping variable. DIF dissection essentially has the same classifications as STAND and smoothed STAND, with the only difference being that it allows the use of either single or multiple grouping variables. Instead of a series of traditional one-way DIF analyses contrasting each focal group with a reference group, DIF dissection allows for the study of interactions among the DIF analysis variables. For example, instead of simply analyzing gender DIF or ethnicity/race DIF separately, they are analyzed simultaneously.

The SIBTEST, kernel smoothed SIBTEST (Douglas, Stout, & DiBello, 1996), and MULTISIB (Stout, Li, Nandakumar, & Bolt, 1997) methods use similar classifications. All three methods use a latent-score definition of null DIF with binary scored items. SIBTEST and kernel smoothed SIBTEST both employ a single grouping variable, while MULTISIB uses a multivariate matching variable. In practice, however, SIBTEST uses an observed score for matching purposes.

4.2 Nonparametric Odds Ratio Methods

Several extensions that have been made to the MH procedure will be discussed in this section. Besides the MH method, the Cochran-Mantel-Haenszel method (CMH; Meyer, Huynh, & Seaman, 2004), the Liu-Agresti estimator (Penfield & Algina, 2003) and Cox's β (Camilli & Congdon, 1999) will be discussed. All the methods in this section share the characteristic of estimating odds ratios using nonparametric techniques. Additionally, all these methods fit the same DIF classification criteria, except for MH, which is only used with binary items. CMH, the Liu-Agresti estimator, and Cox's β can study either binary or polytomous items. These classifications are presented in their entirety in Appendix B.

MH is one of the most commonly used DIF methods, and a description of extensions to MH-related methods will add an important component to our review of recent developments in DIF. The first MH extension used exact, instead of asymptotic, statistics (Parshall & Miller, 1995) to analyze DIF for small samples. Another extension aimed at reducing Type I error and increasing power used MH methods for multiple subgroup comparisons (Penfield, 2001). Additionally, Zwick et al. (1997, 1999, 2000) along with Sinharay, Dorans, Grant, and Blew (in press) provided an enhancement to MH DIF analysis using Bayesian approaches.

An index to measure the variance of differential test functioning (DTF) was proposed by Camilli and Penfield (1997) and then extended to mixed-format tests by Penfield and Algina

(2006). Based on DTF, Penfield and Algina (2006) formulated generalized DIF effect variance estimators to provide indices for evaluating DIF effects across items in a test and translated these into the popular $\log(\alpha_{MH})$ classification scheme (Holland & Thayer, 1988). With this approach, the magnitude of DIF can be evaluated for both the test and its items, with the items being binary, polytomous, or a combination of the two. The DTF approach of Penfield and Algina (2006) is closely related to the Liu-Agresti estimator and Cox's β methods. Another MH-related advancement is the use of the Mantel test of linear association for analyzing polytomous items (Zwick et al., 1993).

In prior research, the CMH procedure was used for analyzing DIF among both binary and polytomous items (Dorans et al., 1992; Holland & Thayer, 1988; Zwick & Thayer, 1996; Zwick, Thayer, & Mazzeo, 1997). However, further research showed that evaluations of DIF by comparing CMH to exact nonparametric approaches (e.g., Wilcoxon Rank Sum Test and van der Waerden Normal Scores test) on both statistical and practical significance, provided evidence that CMH is useful for evaluating polytomous item DIF for small samples (Meyer et al., 2004). The Liu-Agresti estimator estimates common odds ratios and is an interesting generalization of MH in that it employs an effect size estimator and is flexible enough to allow for tests of significance, Bayesian analyses of DIF, and variance-based estimators of DTF. Cox's β was shown to be particularly useful with the partial credit model, but as noted above, it is also useful with binary scored items.

4.3 Generalized Linear Model Methods

The next group of methods is labeled generalized linear model (GLM) methods, since these methods model data that is linearly based on assumed probability distributions (see McCullagh & Nelder, 1989). The GLM methods are of three types: mixture, hierarchical, and logistic. There are five methods altogether, including the following: (a) logistic regression (LR) models, (b) logistic mixed models, (c) mixture models, (d) hierarchical generalized linear models (HGLM), and (e) hierarchical logistic regression (HLR) models. All of the aforementioned methods except for LR have appeared in the literature since 2002.

Each of these methods has interesting characteristics and is worthy of further study. LR was extended to polytomous items (French & Miller, 1996; Rogers & Swaminathan, 1993) and ordered response items (Kristjansson, Aylesworth, McDowell, & Zumbo, 2005) after it was first

proposed by Swaminathan and Rogers (1990) for use with binary scored items. Logistic mixed models (Van den Noortgate & De Boeck, 2005) are based on the assumption that item-level DIF can be modeled using a random effects model. Mixture IRT models provide a basis for understanding the causes of DIF through exploratory mixture model analysis, which is used to define the primary dimension(s) that contribute to DIF as well as the basis for follow-up studies to evaluate examinee characteristics associated with the defined dimensions (Cohen & Bolt, 2005). A further look at mixture models was provided in the mixture distribution conceptualization of de Ayala et al., (2002). Both HGLM (Cheong, 2006; Williams & Beretvas, 2006) and HLR (Swanson, Clauser, Case, Nungster, & Featherman, 2002) take advantage of hierarchical structure in the data and allow simultaneous modeling of additional relevant factors and variables.

The only characteristic that GLM methods have in common is the use of a univariate matching variable—except for LR and HLR, which can use either univariate or multivariate matching. Mixture models and logistic mixed models have a latent-based null-DIF definition and are used with binary scored items. The former can employ both single and multiple grouping variables, while the latter can use only a single grouping variable. HLR is used with binary scored items and either single or multiple grouping variables, and it has a latent score-based definition of null DIF. LR uses only a single grouping variable and a binary or polytomous studied item score. There are two opinions on the definition of null DIF for LR. Shealy and Stout (1993) classify it as a latent-score method, while Potenza and Dorans (1995) classify it as an observed-score method.

HGLM is somewhat different from the other GLM methods. It can be used with both binary or polytomous items, and it is able to employ either single or multiple grouping variables. The classifications for these methods are presented in Appendix C.

4.4 IRT-Based Methods

The last set of methodological developments in DIF has a strong IRT basis. The first method is DFIT, which is an IRT-based framework for assessing differential item functioning of items and tests (Oshima et al., 1997; Oshima, Raju, Flowers, & Slinde, 1998; Raju, van der Linden, & Fleer, 1995). It estimates the expected between-group squared difference in true scores after conditioning on ability and can be used with binary or polytomous data (Bolt, 2002; Flowers, Oshima, & Raju, 1999).

The next method classified is TestGraf. This is a graphical DIF method with kernel smoothing for estimating the conditional probability of correct answers related to proficiency estimates (Bolt & Gierl, 2006; Ramsay, 2000). One part of TestGraf is called TESTCOMP, and it graphically compares the item response functions of reference and focal groups in a DIF analysis. Scrams and McLeod (2000) also formulated a graphical DIF analysis method, which has the big advantage of being able to detect DIF at specific points along the ability scale where subgroup performance differs. The multiple indicator, multiple causes (MIMIC) confirmatory factor analysis model provided another new approach to DIF analysis (Muthén, 2002). It is linked to both the normal ogive IRT model and structural equation modeling.

The last three approaches are the Lagrangian multiplier tests (Glas, 1999), random coefficient multinomial logits (RCML; Moore, 1996) and McDonald's DIF approach (1999). Lagrangian multiplier tests most closely resemble the likelihood ratio test and the Wald test. Their biggest advantage is the flexibility to conduct DIF analyses with both binary and polytomous IRT models. RCML is versatile and useful for detecting DIF based on many different types of Rasch models. Lastly, McDonald's approach is flexible and analyzes DIF using a factor analytic model that mimics IRT models.

All the IRT-based methods use a latent score in their definition of null DIF. Additionally, all the methods are capable of being used with either binary or polytomous items, with the only exception being the MIMIC model, which is used to study only binary items. TestGraf, the Scrams-McLeod method, and the MIMIC model can use either single or multiple grouping variables, while the remaining methods can use only single grouping, with the exception of McDonald's method, for which it was unclear whether it is also capable of using multiple grouping variables. The TestGraf and Scrams-McLeod methods use only a univariate matching variable. The Lagrangian multiplier and RCML methods can use a univariate matching variable, but it is unclear whether these two methods can also use multivariate matching. However, DFIT, the MIMIC model, and McDonald's method can use either univariate or multivariate matching. A summary of DIF classifications for IRT-based methods is provided in Appendix D.

5. Classification Using Statistical Criteria

Statistical criteria that are used for evaluating DIF results or the quality and accuracy of DIF findings are presented in this section. The statistical criteria that will be used for classifying DIF methods are (a) the link to test theory, (b) the existence of an interpretable measure of the

amount of DIF, (c) the existence of a standard error estimate, (d) the existence of a test of significance, and (e) the manner in which the item or matching variable is estimated. These criteria are helpful because they allow practitioners to gauge the usefulness of each method from a theoretical perspective. Appendixes E to H present the statistical criteria used for evaluating these DIF methods.

5.1 Link to Test Theory

Link to test theory describes whether the DIF method is formulated based on one of the two measurement paradigms (i.e., classical test theory or IRT) or uses neither of the two. All the IRT-based methods along with logistic mixed models, mixture models, Cox's β, HLR, and HGLM have an IRT basis, while all the SIBTEST-based methods have both an IRT and classical test theory (CTT) basis. MH, CMH, and the STAND-based methods have a CTT basis, while the Liu-Agresti estimator is not linked to test theory. Although the LR procedure was not derived from test theory, some might consider it to be related to test theory, since the logistic regression function and the linear model in LR both have similar forms in IRT and CTT.

5.2 Interpretable Measure of Amount of DIF

In DIF analyses it is helpful to report the amount of DIF exhibited by an item and not merely whether or not DIF is exhibited. This is akin to reporting effect sizes and is used in ascertaining the practical significance of DIF when it is detected. Expected item score methods all use some form of item score difference. In the case of STAND, it is a standardized difference in the proportion correct between reference and focal group on the studied item, while for SIBTEST the difference between reference and focal groups at each score point is weighted by the proportion of individuals in the focal group who have that score. All the nonparametric odds ratio methods use some form of odds ratio to measure the practical significance of DIF when it is detected. For ease of interpretation, these methods often compute the logarithm of these odds ratios with special formulations such as the MH delta and Liu-Agresti common odds ratio being used for MH and the Liu-Agresti estimators, respectively. Several effect size measures exist for binary LR: (a) R^2-like indices (see Jodoin & Gierl, 2001 and Zumbo, 1999), (b) log odds ratio, and (c) standardized proportion difference correct indices (see Monahan, McHorney, Stump, & Perkins, 2007). Similar measures could be derived for polytomous LR. R^2-like and log odds estimates are used for HLR, while log odds ratios are used for the logistic mixed model.

Mixture models and HGLM use likelihood and log odds ratios, respectively. For RCML, it was unclear whether an interpretable measure of DIF exists, but it appears as though logits could be used for this purpose. Interpretable measures of the amount of DIF were not explicitly described for the Lagrangian multiplier tests and McDonald's method, while DFIT uses compensatory (CDIF) and noncompensatory (NCDIF) DIF statistics for quantifying the amount of DIF. TestGraf measures the amount of DIF using a root mean square average difference between each focal group and the reference group for individual items (Gierl & Bolt, 2001). Lastly, the Scrams-McLeod method uses the MH statistic to measure the amount of DIF.

5.3 Standard Error Estimate

Reporting standard errors helps in assessing the amount of random variability associated with DIF estimates. Generally, expected item score methods, nonparametric odds ratio methods, and GLM methods (with the exception of mixture models) have standard error estimates. All the IRT-based methods have standard error estimates except for the CDIF component of DFIT.

5.4 Test of Significance

SIBTEST methods use a test of significance based on the ratio of $\hat{\beta}$ (a parameter estimate specifying the amount of DIF) to its standard error. MH, CMH, LR, mixture models, and DFIT use a chi-square test of significance. DFIT can also use a *t* test of significance, but these significance tests do not apply to CDIF. HLR, HGLM, and the MIMIC model detect DIF by testing the significance of model coefficients, and RCML uses Hotelling's T. The Liu-Agresti estimator uses a cumulative common odds ratio index, Cox's β uses a β statistic, logistic mixed models use the Wald test, while Lagrangian multiplier tests have a test of significance related to the difference between the observed and expected number of persons in the focal group scoring in a particular category (i.e., correct/incorrect or polytomous category) for each item.

5.5 Estimation of Item/Matching Variable Relationship

Some DIF methods use a functional form for modeling the relationship between item score and matching variable; these are referred to as parametric. Nonparametric estimation may be preferable because it does not make any model-based assumptions about the form of the item/ability regression. All the nonparametric odds ratio and expected item score methods

employ nonparametric estimation. All the IRT-based methods employ parametric estimation except for TestGraf. Similarly, all the GLM methods employ parametric models.

6. Classification Using Practical Criteria

To be used in practice, a DIF method must be efficient as well as easy to use. Sequential, iterative evaluation at each step is a luxury that cannot be afforded in practice. The practical criteria used for classifying the DIF methods discussed in this paper are (a) procedural requirements, (b) computational intensity, and (c) cost. A summary of classifications based on practical criteria is also presented in Appendixes E to H.

6.1 Procedural Requirements

More complex models often require more time to manipulate and process variables. Therefore, preferred methods have simple underlying models with simple procedural requirements. All the nonparametric odds ratio methods, STAND methods, LR, logistic mixed models, the Scrams-McLeod method, TestGraf, RCML, SIBTEST methods, and McDonald's method are simple, since they do not require difficult manipulation of data or variables, repeated analyses, or iterations to ensure accuracy or model fit. The computer software is also easy to use. The procedural requirements of the HLR, Lagrangian multiplier tests, and HGLM are moderately demanding, since an appreciable amount of data or software manipulation is required.

The DFIT, mixture model, and MIMIC models are viewed as labor intensive. Most notably, mixture models use Markov chain Monte Carlo (MCMC) methods, which can sometimes take a considerable amount of time for solutions to be found (due to computational demands). Moreover, the MIMIC and DFIT methods sometimes require multiple evaluations and the use of various combinations of variables before analyses can be considered complete.

6.2 Computational Intensity

The amount of time it takes to manipulate data and complete DIF analysis computations is of critical importance in practice. Specifically, speed and efficiency are valued under stringent analysis and reporting schedules. Additionally, the computational efficiency of DIF methods adds to cost savings. Nonparametric methods tend to have lower computational intensity. The GLM methods are generally more computationally intensive, with mixture models being the most intensive because they use MCMC methods. However, with increased computing power,

computational intensity is not as big an issue as it has been in the past; however, convergence may be an issue with MCMC methods. The computational requirements of MULTISIB are not clear.

6.3 Cost

The degree to which a DIF method can be used with ease in practice is largely a function of the human resources needed to evaluate results and the extent to which these results are easy to interpret, along with the aforementioned practical criteria. Hence, the expected item scores methods, TestGraf, RCML, McDonald's method, and LR are likely to be inexpensive. The cost of the nonparametric odds ratio and Scrams-McLeod approach methods is variable depending on the number of sets of odds ratios being studied. MULTISIB's cost is also variable, since dimensionality analyses are not always easy or straightforward. Methodological complexities (e.g., data manipulation, iterative variable, and fit evaluations) led to the classification of HGLM, HLR, the MIMIC model, and Lagrangian multiplier tests as likely to be expensive. Logistic mixed models, mixture models, and DFIT are also likely to be expensive due to the complexity of parameter estimation, high labor intensity, and the need for iterative evaluations of model fit.

7. Discussion and Conclusions

The purpose of this paper was to evaluate DIF methods that have appeared over the past 15 years in the research literature. Specifically, the focus was on methods that appeared after the publication of the Potenza and Dorans (1995) DIF taxonomy. This paper was motivated by the need to address recent challenges (e.g., small sample sizes) and emerging testing contexts (e.g., the increased numbers of non-native English-speaking examinees and examinees with learning disabilities). Many of the innovations in the literature have failed to address these issues, suggesting a disconnect between the interests of the theoreticians and the needs of the practitioner .Some of the evaluated methods appear promising. There are also other noteworthy extensions to DIF analyses that will be discussed briefly below.

As noted previously, one of the motivations for this review was to find DIF methods that work well with small samples. CMH (Meyer et al., 2004; Parshall & Miller, 1995) has been studied as a potential solution and may be worthy of further investigation, because comparisons of its statistical and practical significance findings to those of exact nonparametric approaches produced comparable results. Other approaches to small sample DIF analysis, for example, Bayesian methods (Sinharay et al., in press) or smoothing techniques (Puhan et al., 2007), have

not succeeded in lowering the minimum sample size threshold. Therefore, continued research in this area is still needed.

Increased recognition of the complexity of tests implies the future importance of multidimensional DIF. This approach has already been illustrated in several studies (Camilli, 1992; Gierl, Bisanz, Bisanz, & Boughton, 2003; Walker & Beretvas, 2001). Even so, there are only a few promising approaches (Roussos & Stout, 1996; Stout et al., 1997), and they are unlikely to be used operationally. Therefore, more research should be conducted to find methods that can be used in practice.

Another promising DIF approach is differential distractor functioning (DDF). This approach entails the analysis of distractor choices among those who answer an item incorrectly, but it cannot be considered a new method since it is based on STAND (Dorans et al., 1992) and log-linear approaches (Green, Crone, & Folk, 1989). DDF could help test developers understand group differences in testing through the analysis of differences in response option choices and provide a means for supporting substantive and qualitative interpretation of DIF analyses. DDF is predicated on the premise that incorrect answer options are differentially attractive to examinees of different backgrounds. It has the potential of providing supporting data and analysis to corroborate or refute proposed reasons why subgroup response differences may or may not be construct relevant, or to determine whether DIF might be attributable to specific features of an item (such as a specific distractor).

Using the DTF approach, the overall impact of DIF effects when combined across the items in a test can be studied for binary, polytomous, and multidimensional tests (see McDonald, 1999; Raju et al., 1995). Out of all the observed score matching methods we evaluated, the Liu-Agresti estimator was the only one that illustrated the use of the DTF framework. Several latent score matching methods that were evaluated are capable of performing DTF analyses (e.g., SIBTEST, DFIT, and McDonald's method). Aside from score equity assessment (Dorans, 2004), the only other approach to detecting the differential prediction of a test is through DTF analysis.

Given the wide variety of new DIF methods, it is likely that their efficiency at detecting problematic items will vary. Comparisons of these new methods would be an important next step in ascertaining their suitability for practical use. This type of evaluation could be completed with the participation of test developers and content experts so that the efficiency of each method is thoroughly evaluated. Additionally, extensive and rigorous study of these DIF methods would be

needed in order to determine minimum and maximum sample sizes required for drawing accurate inferences.

Lastly, during our review we found an interesting paradigmatic debate that is related to the practical applicability of these new DIF methods. This debate relates to the Holland (1994) and Wainer (1993) discussion of the impact and merits of viewing tests according to the themes of measurement versus contest. When a test is viewed as a measurement, the focus is on its measurement properties such as reliability and validity. This view is consistent with modeling examinee ability in terms of underlying latent traits. In the contest view, however, the emphasis is on fair play and is consistent with the simple "number correct" score approach, which is easily understood by the examinee and easy to compute. Hence, from a contest perspective, and given the fact that DIF was designed to address the contest aspects of testing, procedures that employ a null DIF definition that conditions on the observed score of interest, namely the reported test score, would be preferred over those that do not condition or match on the reported score. Therefore, if the contest view were to be adopted, half the methods in Appendixes A and C and all the methods in Appendix D would be eliminated from consideration.

In this paper, several DIF, statistical and practical criteria were proposed for evaluating the efficiency and appropriateness of DIF methods for practical use. These criteria will be helpful to practitioners as they appraise the applicability of new DIF methods in their testing programs. Moreover, this paper fosters a continual updating of knowledge about DIF that in turn will encourage and ensure the enhancement of test quality and equity in practice.

References

Bolt, D. M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education, 15*(2)*,* 113–141.

Bolt, D. M., & Gierl, M. J. (2006). Testing features of graphical DIF: Application of a regression correction to three nonparametric statistical tests*. Journal of Educational Measurement*, *43,* 313–333.

Camilli, G. (1992). A conceptual analysis of differential item functioning in terms of a multidimensional item response model. *Applied Psychological Measurement, 16*(2)*,* 129–147.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 221–256). Westport, CT: Praeger.

Camilli, G., & Congdon, P. (1999). Application of a method of estimating DIF for polytomous test items. *Journal of Educational and Behavioral Statistics*, *24*(4), 323–341.

Camilli, G., & Penfield, D. A. (1997). Variance estimation for differential test functioning based on Mantel-Haenszel statistics. *Journal of Educational Measurement, 34*(2)*,* 123–139.

Chang, H.-H, Mazzeo, J., & Roussos, L. A. (1993, April). *Extension of Shealy-Stout's DIF procedures to polytomously scored items.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

Chang, H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33*(3), 333–353.

Cheong, Y. F. (2006). Analysis of school context effects on differential item functioning using hierarchical generalized linear models**.** *International Journal of Testing, 6*(1), 57–79.

Clauser, B, E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*(1), 31–44.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement, 42*(2)**,** 133–148.

de Ayala, R. J., Kim, S.-H., Stapleton, L. M., & Dayton, C. M. (2002). Differential item functioning: A mixture distribution conceptualization. *International Journal of Testing, 2*(3 & 4), 243–276.

Dorans, N. J. (2004). Using subpopulation invariance to assess test score equity. *Journal of Educational Measurement, 41*(1), 43–68.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*(4)*,* 355–368.

Dorans, N. J., & Schmitt, A. P. (1993). Constructed response and differential item functioning: A pragmatic perspective. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 135–165). Hillsdale, NJ: Lawrence Erlbaum.

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.

Douglas, J., Roussos, L., & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement, 33*(4)*,* 465–484.

Douglas, J. A., Stout, W., & DiBello, L. V. (1996). A kernel-smoothed version of SIBTEST with application to local DIF inference and function estimation. *Journal of Educational Measurement, 21*(4)*,* 333–363.

Flowers, C. P., Oshima, T. C., & Raju, N. S. (1999). A description and demonstration of the polytomous-DFIT framework. *Applied Psychological Measurement, 23*(4), 309–326.

French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement, 33*(3), 315–332.

Gierl, M., Bisanz, J., Bisanz, G., & Boughton, K. (2003). Identifying content and cognitive skills that produce gender differences in mathematics: A demonstration of the multidimensionality-based DIF analysis paradigm. *Journal of Educational Measurement, 40*(4), 281–306.

Gierl, M., Bisanz, J., Bisanz, G., Boughton, K., & Khaliq, S. (2001). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests. *Educational Measurement: Issues and Practice, 20*(2)*,* 26–36.

Gierl, M. J., & Bolt, D. M. (2001). Illustrating the use of nonparametric regression to assess differential item and bundle functioning among multiple groups. *International Journal of Testing, 1*(3 & 4)*,* 249–270.

Glas, C.A.W. (1999). Detection of differential item functioning using Lagrange multiplier tests. *Statistica Sinica*, *8*(1), 647–667.

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.

Haladyna. T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17–26.

Holland, P. W. (1994). Measurements or contests? Comments on Zwick, Bond, and Allen/Donoghue. *1994 proceedings of section on social statistics, American Statistical Association*, 27–29. Alexandria, VA: American Statistical Association.

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

Individuals with Disabilities Education Act of 1991, 20 U.S.C. § 1400 *et seq*. (1991).

Individuals with Disabilities Education Act of 1997, 20 U.S.C. § 1412(a) (17)(A). (1997).

Jodoin, M. G., & Gierl, M.J. (2001). Evaluating Type I error and power rates using an effect size measure with logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*(4), 329–349.

Kristjansson, E., Aylesworth, R., McDowell, I, & Zumbo, B. D. (2005). A comparison of four methods for detecting DIF in ordered response items. *Educational and Psychological Measurement*, *65*(6), 935–953.

Li, H., & Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika, 61*(4), 647–677.

Lyu, F., Dorans, N., & Ramsay, J. O. (1995, April). *Smoothed standardization assessment of testlet level DIF on a math free-response item* (ETS Research Rep. No. RR-95-38). Princeton, NJ: ETS.

Mantel, N. (1963). Chi-square tests with one degree of freedom; Extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58*(303), 690–700.

Mantel, N., & Haenszel, W. M. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*(4)*,* 719–748.

Mazor, K. M., Kanjee, A., & Clauser, B. E. (1995). Using logistic regression and the Mantel-Haenszel with multiple ability estimates to detect differential item functioning. *Journal of Educational Measurement, 32*(2), 131–144.

McCullagh, P. & Nelder, J.A. (1989). *Generalized linear models*. Chapman and Hall: London.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Hillsdale, NJ: Erlbaum.

Meyer, J. P., Huynh, H., & Seaman, M. A. (2004). Exact small-sample differential item functioning methods for polytomous items with illustration based on an attitude survey. *Journal of Educational Measurement*, *41*(4), 331–344.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*(4), 297–334.

Monahan, P. O., McHorney, C. A., Stump, T. E., & Perkins, A. J. (2007). Odds ratio, delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression. *Journal of Educational and Behavioral Statistics, 32*(1), 92–109

Moore, S. (1996). Estimating differential item functioning in the polytomous case with the random coefficient multinomial logits (RCML) model. In G. Englehard & M. Wilson (Eds.), *Objective measurement III: Theory into practice* (pp. 219–238). Norwood, NJ: Ablex.

Muraki, E. (1993, April). *Implementing item parameter drift and bias in polytomous item response models*. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika*, *29*(1), 81–117.

Oshima, T. C., Raju, N. S., & Flowers, C. P. (1997). Development and demonstration of multidimensional IRT-based internal measures of differential functioning of items and tests**.** *Journal of Educational Measurement, 34*(3), 253–272.

Oshima, T. C., Raju, N. S., Flowers, C. P., & Slinde, J. A. (1998). Differential bundle functioning using the DFIT framework: Procedures for identifying possible sources of differential functioning**.** *Applied Measurement in Education, 11*(4), 353–369.

Parshall, C. G., & Miller, T. R. (1995). Exact versus asymptotic Mantel-Haenszel DIF statistics: A comparison of performance under small-sample conditions. *Journal of Educational Measurement, 32*(3), 302–316.

Penfield, R. D. (2001). Assessing differential item functioning among multiple groups: A comparison of three Mantel–Haenszel procedures. *Applied Measurement in Education, 14*(3)*,* 235–259.

Penfield, R. D., & Algina, J. (2003). Applying the Liu-Agresti estimator of the cumulative common odds ratio to DIF detection in polytomous items. *Journal of Educational Measurement*, *40*(4), 353–370.

Penfield, R. D., & Algina, J. (2006). A generalized DIF effect variance estimator for measuring unsigned differential test functioning in mixed format tests. *Journal of Educational Measurement 43*(4), 295–312.

Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement, 19*(1), 23–37.

Puhan, G., Moses, T. P., Yu, L., & Dorans, N. J. (2007). *Small-sample DIF estimation using log-linear smoothing: A SIBTEST application* (ETS Research Rep. No. RR-07-10). Princeton, NJ: ETS.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based measures of differential item functioning of items and tests. *Applied Psychological Measurement, 19*(4)*,* 353–368.

Ramsay, J. O. (2000). TestGraf: A program for the graphical analysis of multiple-choice test and questionnaire data [Computer program and manual]. Retrieved June 2, 2008, from http://www.psych.mcgill.ca/faculty/ramsay/ramsay.html.

Rogers, H. J., & Swaminathan, H. (1993, April). *Differential item functioning procedures for non-dichotomous responses*. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*(4), 355–371.

Schmitt, A. P., Curley, W. E., Bleistein, C. A., & Dorans, N. J. (1988, April). *Experimental evaluation of language and interest factors related to differential item functioning for*

Hispanic examinees on the SAT-Verbal. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Schmitt, A. P, Holland, P. W., & Dorans, N. J. (1993). Evaluating hypotheses about differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 281–315). Hillsdale, NJ: Lawrence Erlbaum.

Scrams, D. J., & McLeod, L. D. (2000). An expected response function approach to graphical differential item functioning. *Journal of Educational Measurement, 37*(3), 263–280.

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

Sinharay, S., Dorans, N. J., Grant, M., & Blew, E. (in press). Using past data to enhance small-sample DIF estimation: A Bayesian approach. *Journal of Educational and Behavioral Statistics.*

Sinharay, S., Liang, L., & Dorans, N. J. (2008, March). *English proficiency and its effects on fairness assessment*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Stone, E. A., Cook, L. L., Cline, F., & Cahalan-Laitusis, C. (2007, April). *Using differential item functioning to investigate the impact of testing accommodations on an English language arts assessment for students who are blind and visually impaired*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Stout, W., Li, H.-H., Nandakumar, R., & Bolt, D. (1997). MULTISIB: A procedure to investigate DIF when a test is intentionally two-dimensional. *Applied Psychological Measurement, 21*(3), 195–215.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361–370.

Swanson, D. B., Clauser, B. E., Case, S. M., Nungster, R. J., & Featherman, C. (2002). Analysis of differential item functioning (DIF) using hierarchical logistic regression models. *Journal of Educational and Behavioral Statistics, 27*(1), 53–75.

Van den Noortgate, W., & De Boeck, P. (2005) Assessing and explaining differential item functioning using logistic mixed models. *Journal of Educational and Behavioral Statistics*, *30*(4), 443–464.

Wainer, H. (1983). Pyramid power: Searching for an error in test. scoring with 830,000 helpers. *The American Statistician, 37*, 87–81.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. *Applied Measurement in Education. 8*(2), 157–186.

Walker, C. M., & Beretvas, S. N. (2001). An empirical investigation demonstrating the multidimensional DIF paradigm: A cognitive explanation for DIF. *Journal of Educational Measurement, 38*(2), 147–163.

Welch, C., & Hoover, H. D. (1993). Procedures for extending item bias techniques to polytomously scored items. *Applied Measurement in Education, 6*, 1–19.

Williams, N. J., & Beretvas, S. N. (2006). DIF identification using HGLM for polytomous items. *Applied Psychological Measurement, 30*(1), 22–42.

Zhang, Y., Dorans, N., & Matthews-Lopez, J. (2005). *Using DIF dissection method to assess effects of item deletion* (ETS Research Rep. No. RR-05-23). Princeton, NJ: ETS.

Zieky, M. J. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning*. (pp. 337–47). Hillsdale, NJ: Lawrence Erlbaum.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF)*. Ottawa, Ontario, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*(3), 233–251.

Zwick R., & Thayer, D. T. (1996). Evaluating the magnitude of differential item functioning in polytomous items. *Journal of Educational Statistics, 21*(3), 187–201.

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement, 36*(1), 1–28.

Zwick, R., & Thayer, D. T., & Lewis, C. (1997). *An investigation of the validity of an empirical Bayes approach to Mantel-Haenszel DIF analysis* (ETS Research Rep. No. RR-97-21). Princeton, NJ: ETS.

Zwick, R., Thayer, D. T., & Lewis, C. (2000). Using loss functions for DIF detection: An empirical Bayes approach. *Journal of Educational and Behavioral Statistics, 25*(1), 225–247.

Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). Descriptive and inferential procedures for assessing differential item functioning in polytomous items. *Applied Measurement in Education, 10*(4)*,* 321–344.

Notes

[1] This paper uses STAND as the same acronym for what has been called STD or STND in the past.

[2] SIBTEST is a latent trait approach in theory. In practice it uses estimated true scores, which are group-specific transformed observed scores, as a matching variable. Hence it is difficult to classify.

Appendix A

Classification of Expected Item Score Methods

Category	Description	Method					
		STAND	Smoothed STAND	DIF dissection	SIBTEST	Kernel smoothed SIBTEST	MULTISIB
Source		Dorans & Kulick (1986)	Lyu, Dorans, & Ramsay (1995)	Zhang, Dorans, & Matthews-Lopez (2005)	Shealy & Stout (1993)	Douglas, Stout, & DiBello (1996)	Stout, Li, Nandakumar, & Bolt (1997)
Null DIF[a]	P(Y|O,G) = P(Y|O) [O] vs. P(Y|L,G) = P(Y|L) [L]	O	O	O	L	L	L
Studied item score	Binary (B) vs. polytomous (P)	B	B	B	B	B	B
Matching variable	Univariate (U) vs. multivariate (M)	U	U	U	U	U	M
Grouping variable	Single (S) vs. multiple (M)	S	S	S/M	S	S	S

Notes. DIF = differential item functioning. STAND = standardization.

[a] Where Y = observed score (univariate/multivariate) provided by a measuring instrument as a random variable, L = latent/unobserved variable for which Y is the intended observed indicator (can be univariate/multivariate), O = observed total test score (univariate/multivariate), which can serve as a stratifying variable when examining DIF, G = grouping variable (univariate/multivariate), typically of demographic information (e.g., race, gender).

Appendix B

Classification of Nonparametric Odds Ratio Methods

Category	Description	Method			
		Mantel-Haenszel	Cochran-Mantel-Haenszel	Liu-Agresti estimator	Cox's β
Source		Holland & Thayer (1988)	Meyer, Huynh, & Seaman (2004); Parshall & Miller (1995)	Penfield & Algina (2003)	Camilli & Congdon (1999)
Null DIF[a]	P(Y|O,G) = P(Y|O) [O] vs. P(Y|L,G) = P(Y|L) [L]	O	O	O	O
Studied item score	Binary (B) vs. polytomous (P)	B	B/P	B/P	B/P
Matching variable	Univariate (U) vs. multivariate (M)	U	U	U	U
Grouping variable	Single (S) vs. multiple (M)	S	S	S	S

Notes. DIF = differential item functioning.

[a] Where Y = observed score (univariate/multivariate) provided by a measuring instrument as a random variable, L = latent/unobserved variable for which Y is the intended observed indicator (can be univariate/multivariate), O = observed total test score (univariate/multivariate), which can serve as a stratifying variable when examining DIF, G = grouping variable (univariate/multivariate), typically of demographic information (e.g., race, gender).

Appendix C

Classification of Generalized Linear Model Methods

Category	Description	Method				
		Logistic regression	Hierarchical logistic regression	Logistic mixed model	Mixture model	HGLM
Source		Swaminathan & Rogers (1990); French & Miller (1996); Rogers & Swaminathan (1993)	Swanson, Clauser, Case, Nungster, & Featherman (2002)	Van den Noortgate & De Boeck (2005)	Cohen & Bolt (2005)	Cheong (2006); Williams & Beretvas (2006)
Null DIF[a]	P(Y|O,G) = P(Y|O) [O] vs. P(Y|L,G) = P(Y|L) [L]	O/L	L	L	L	L
Studied item score	Binary (B) vs. polytomous (P)	B/P	B	B	B	B/P
Matching variable	Univariate (U) vs. multivariate (M)	U/M	U/M	U	U	U
Grouping variable	Single (S) vs. multiple (M)	S	S/M	S	S/M	S/M

Note. DIF = differential item functioning; HGLM = hierarchical generalized linear model.

[a] Where Y = observed score (univariate/multivariate) provided by a measuring instrument as a random variable, L = latent/unobserved variable for which Y is the intended observed indicator (can be univariate/multivariate), O = observed total test score (univariate/multivariate), which can serve as a stratifying variable when examining DIF, G = grouping variable (univariate/multivariate), typically of demographic information (e.g., race, gender).

Appendix D

Classification of IRT-Based Methods

Category	Description	Method						
		DFIT	TestGraf	Scrams-McLeod	MIMIC model	Lagrangian multiplier tests	RCML	McDonald's
Source		Raju, van der Linden, & Fleer (1995); Oshima, Raju & Flowers (1997); Flowers, Oshima, & Raju (1999)	Ramsay (2000); Gierl & Bolt (2001)	Scrams & McLeod, (2000)	Muthén (2002)	Glas (1999)	Moore (1996)	McDonald (1999)
Null DIF[a]	P(Y|O,G) = P(Y|O) [O] vs. P(Y|L,G) = P(Y|L) [L]	L	L	L	L	L	L	L
Studied item score	Binary (B) vs. polytomous (P)	B/P	B/P	B/P	B	B/P	B/P	B/P
Matching variable	Univariate (U) vs. multivariate (M)	U/M	U	U	U/M	U/?	U/?	U/M
Grouping variable	Single (S) vs. multiple (M)	S	S/M	S/M	S/M	S	S	S/?

Note. DFIT = differential functioning of items and test; DIF = differential item functioning; MIMIC = multiple indicator, multiple causes; RCML = random coefficient multinomial logits.

[a] Where Y = observed score (univariate/multivariate) provided by a measuring instrument as a random variable, L = latent/unobserved variable for which Y is the intended observed indicator (can be univariate/multivariate), O = observed total test score (univariate/multivariate), which can serve as a stratifying variable when examining DIF, G = grouping variable (univariate/multivariate), typically of demographic information (e.g., race, gender).

Appendix E

Evaluation of Expected Item Score Methods Based on Statistical and Practical Criteria

Criteria	Method					
	STAND	Smoothed STAND	DIF dissection	SIBTEST	Kernel smoothed SIBTEST	MULTISIB
Statistical						
Link to test theory	CTT	CTT	CTT	IRT & CTT	IRT & CTT	MIRT & CTT
Interpretable measure of amount of DIF	Standardized expected item score measure in focal group metric	Standardized expected item score measure in focal group metric	Standardized expected item score measure in focal group metric	Weighted mean difference	Weighted mean difference	Weighted mean difference
Standard error estimate	Yes	Yes	Yes	Yes	Yes	Yes
Test of significance	None	None	None	SIB test statistic	SIB test statistic	SIB test statistic
Estimation of item/matching variable relationship	Nonparametric	Nonparametric	Nonparametric	Nonparametric	Nonparametric	Nonparametric
Practical						
Procedural requirements	Simple	Simple	Simple	Simple	Simple	Simple
Computational intensity	Low	Low	Low	Low	Low	?
Cost	Inexpensive	Inexpensive	Inexpensive	Inexpensive	Variable – based on smoothing efficiency	Variable – depending on simplicity of dimensionality analyses

Note. CTT = classical test theory, DIF = differential item functioning, IRT = item response theory, MIRT = multidimensional item response theory, STAND = standardization.

Appendix F

Evaluation of Nonparametric Odds Ratio Methods Based on Statistical and Practical Criteria

	Method			
Criteria	Mantel-Haenszel	Cochran-Mantel-Haenszel	Liu-Agresti estimator	Cox's β
Statistical				
Link to test theory	CTT	CTT	None	IRT
Interpretable measure of amount of DIF	MH log odds ratio estimate	A set of odds ratios	Liu-Agresti cumulative common odds ratio	Log odds ratios
Standard error estimate	Yes	Yes	Yes	Yes
Index or test of significance based	MH chi-square test of significance	Chi-square test of significance	Liu-Agresti cumulative common odds ratio	Cox's β test statistic
Estimation of item/matching variable relationship	Nonparametric	Nonparametric	Nonparametric	Nonparametric
Practical				
Procedural requirements	Simple	Simple	Simple	Simple
Computational intensity	Low	Low	Low	Low
Cost	Variable – depends on how many sets of odds ratios are studied	Variable – depends on how many sets of odds ratios are studied	Variable – depends on how many sets of odds ratios are studied	Variable – depends on how many sets of odds ratios are studied

Note. CTT = classical test theory, DIF = differential item functioning, IRT = item response theory, MH = Mantel-Haenszel.

Appendix G

Evaluation of Generalized Linear Model Methods Based on Statistical and Practical Criteria

Criteria	Method				
	Logistic regression	Hierarchical logistic regression	Logistic mixed model	Mixture model	HGLM
Statistical					
Link to test theory	None	IRT	IRT	IRT	IRT
Interpretable measure of amount of DIF	P-DIF, R^2-like indices and log odds ratios can be used	Effect size estimate based on log odds ratio or R^2-like index based on change in variance components	Log odds ratio estimate conditional on latent ability	Likelihood ratio	Log odds ratio
Standard error estimate	Yes	Yes	Yes	Yes	Yes
Index or test of significance based	Chi-square test of significance	Significance test of model coefficients	Wald test of significance	Chi-square test of significance	Chi-square significance test of model coefficients
Estimation of item/matching variable relationship	Parametric – requires fitting a logistic regression model for probability of answering an item correctly given fixed observed score	Parametric – requires fitting a hierarchical linear model given examinees nested within items	Parametric – fits logistic mixed model	Parametric – assumes latent ability and latent class membership	Parametric – assumes data fit a model in which subgroups are nested within items
Practical					
Procedural requirements	Simple	Moderately demanding	Simple	Labor intensive	Moderately demanding
Computational intensity	Low	Low	High	High	Low
Cost	Inexpensive	Probably expensive – due to methodological complexities	Probably expensive – due to complexity of parameter estimation	Expensive –due to computational requirements of MCMC methods	Probably expensive – because of methodological complexities

Note. DIF = differential item functioning, HGLM = hierarchical generalized linear model, IRT = item response theory, MCMC = Markov chain Monte Carlo.

Appendix H

Evaluation of IRT-Based Methods Based on Statistical and Practical Criteria

	Method						
Criteria	DFIT	TestGraf	Scrams-McLeod	MIMIC model	Lagrangian multiplier tests	RCML	McDonald's
Statistical							
Link to test theory	IRT	IRT	IRT	IRT	IRT	IRT	IRT
Interpretable measure of amount of DIF	Compensatory (CDIF) & noncompensatory (NCDIF) DIF statistics	Root mean square average difference	MH statistic	No	No	In theory	No
Standard error estimate	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Index or test of significance based,	Chi-square & t-tests of significance	None	None	Significance test of model coefficients	Lagrangian multiplier test of significance	Hotelling's T	None
Estimation of item/matching variable relationship	Parametric – data fits an IRT model	Nonparametric	Parametric – data fits an IRT model	Parametric – data fits a CFA model	Parametric – data fits an IRT model	Parametric – data fits an IRT model	Parametric – factor analytic approach
Practical							
Procedural requirements	Labor intensive	Simple	Simple	Labor intensive	Moderately demanding	Simple	Simple
Computational intensity	Low	Low	Low	Low	Low	Low	Low
Cost	Expensive – labor intensive	Inexpensive – since it is simple to understand and interpret	Variable – depends number of sets of odds ratios studied	Probably expensive – because of methodological complexities	Probably expensive – because of methodological complexities	Inexpensive –simple to understand and interpret	Inexpensive –simple to understand and interpret

Note. DIF = differential item functioning, IRT = item response theory, MIMIC = multiple indicator, multiple causes, RCML = random coefficient multinomial logits.