The Aoyama Journal of International Politics, Economics and Communication, No. 97, November 2016

Article

Detecting Differential Item Functioning Using the General Linear Model Framework: A Simulation Study

Katsuya Tasaki[1)]

Abstract

The purpose of this study is to examine the efficacy of a bias-detection method within the general linear model (GLM) framework. I evaluated detection capability of a GLM-based method by comparing it to a confirmatory factor analysis likelihood ratio (CFA-LR) test using simulation data. The GLM method works fairly well for detecting uniform bias under various conditions. Although nonuniform bias detection was somewhat poor under certain conditions, the overall bias-detection pattern of the GLM and CFA-LR methods was quite similar. Across all conditions, the false-positive rates of the GLM were less than the expected value of alpha=.05. Considering its simplicity and flexibility, the GLM method is a powerful and valid alternative for detecting bias when evaluating measurement equivalence.

Key words: DIF, uniform bias, nonuniform bias, MACS, measurement equivalence/invariance, Monte Carlo simulation

Over the past decade, the amount of attention paid to measurement equivalence by the cross-cultural research community has been growing (Byrne & Watkins, 2003; Meade & Lautenschlager, 2004). The establishment of instrument equivalence is imperative for any group comparison,

[1] Professor of Communication Studies, Department of International Communication, Aoyama Gakuin University. Email: tasaki@sipeb.aoyama.ac.jp

but it is particularly critical in cross-cultural comparative studies. Cross-cultural investigation is a unique research process that entails a variety of theoretical and methodological challenges, including: translating assessment instruments into the target language equivalently (Orlando & Marshall, 2002), selecting comparable samples or sub-cultures across countries (Häder & Gabler, 2003) and controlling for culturally conditioned response styles (Kankaraš & Moors, 2011). These challenges can be potential noise or bias for equivalent measures. Cross-cultural researchers cannot readily assume the comparability of test scores in such a way as intracultural survey (Welkenhuysen-Gybels, Billet, & Cambré, 2003). Without confirming that instruments are bias-free and function equally across cultures, survey results may be ambiguous at best and misleading at worst (Steenkamp & Baumgarter, 1998).

For example, since Markus and Kitayama proposed the idea of self-construal in 1991, this concept has been used in many cross-cultural studies. However, some researchers have viewed the validity of self-construal research with skepticism (Levine et al., 2003; Matsumoto, 1999). Based on a meta-analytic review, Levine *et al.* (2003) reported that cultural differences in self-concepts across published articles are weak, inconsistent, or non-existent. They claimed that some cross-cultural studies do not support the self-construal theoretical perspective that people in the US are higher in independent self and lower in interdependent self than their Asian counterparts (e.g. Gudykunst *et al.*, 1996) and that some studies even suggest the existence of cultural differences in self-concepts that are the opposite of those predicted by the theory (e.g. Kleinknecht *et al.*, 1997).

Research findings that are inconsistent with, or even contradictory to, theories may be the result of measurement artifacts. Previous cross-cultural studies rarely evaluated their measurement equivalence of the assessment instruments, as was shown in the case of self-construal research. When assessments do not have measurement equivalence, it is difficult for cross-cultural investigators to meaningfully interpret observed score differences. That is, score differences may reflect true cultural differences between groups, or they may be the result of different relationships between latent and observed variables across the groups that are being compared. Therefore, measurement equivalence is indispensable when comparing and contrasting cultures across demographic groups via test scores, and true cultural differences must be distinguished from measurement artifacts.

Currently, there are two approaches to researching and assessing measurement equivalence: confirmatory factor analysis (CFA) and item response theory (IRT). Both approaches are based on sophisticated statistical models and are the most promising, state-of-the-art methods. However, for researchers who face the "constraints" of real research situations, these approaches are not always ideal. The two methods, particularly the IRT-based method, usually require large samples for efficient parameter estimation. Additionally, statistical software, such as LISREL (Jöreskog & Sörbom, 2002) or MULTILOG (Thissen, 1991), is required to estimate the parameters and to fit the model to the data. Furthermore, conducting an item analysis using these advanced tools is not always a straightforward task.

The present simulation study explores an alternative way to examine measurement equivalence. Using the general linear model (GLM) framework, this study focuses on evaluating measurement equivalence by detecting differential item functioning (DIF). Given its simplicity and practicality, this regression-based DIF detection method is an efficacious alternative that cross-cultural investigators can reasonably substitute for CFA- or IRT-based methods.

Psychometric properties of equivalent measures

In general, measurement equivalence is defined as the invariant operation of test items because they are perceived and interpreted in the same way across groups being compared (Byrne & Watkins, 2003). A more elaborate definition of measurement equivalence accounts for the structural relationship between scales and psychological constructs. That is, measurement equivalence is tenable when the observed item scores and the latent traits that the items measure are identical across the compared groups (Drasgow, 1984). Measurement equivalence is often viewed as an intrinsic property of the instrument. However, it is more practical to view measurement equivalence as the interaction between instrument features and cultural-group characteristics (van de Vijver & Leung, 1997).

"Bias" is a closely related concept because it is a factor that threatens measurement equivalence. Theoretically, equivalence and bias are opposed to one another because observed scores become equivalent when they are unbiased (van de Vijver & Leung, 1997). Therefore, bias can be a "cause" that affects the establishment of equivalence between measures. Unlike reliability indices, such as Cronbach's alpha, which explain random measurement error, bias leads to systematic measurement inaccuracy, which can be replicated over repeated measures (Millsap, 2011). A reliable test is not necessarily unbiased because a highly reliable but biased test will consistently yield inaccurate measurements over repeated measures.

As it is more inclined to measurement artifacts at the item level, "different item functioning" (DIF) is referred to as "item bias." Researchers prefer this term because "bias" has a negative connotation, as it can imply unfairness or prejudice caused by inherent test item flaws (Angoff, 1993). DIF is a value-free term that merely implies something about the statistical findings of different test-item functions across groups (Millsap, 2011). For the purpose of this simulation study, DIF and bias will be used interchangeably to imply different test item performance across cultural groups.

The current psychometric theory of bias detection relies on the matching principle for diagnosing a biased test item (Angoff, 1993). According to this principle, an item is biased when a respondent with the same ability or attribute but in a different cultural group gives a different response to the test item. By matching or conditioning based on the proficiency that the items measure, this approach allows researchers to evaluate the function of the studied item after controlling for group differences in ability or attributes. This matching variable can be either internal or external; an internal matching variable includes the total test score, which is the most widely used criterion, even though it is potentially subject to circularity because the criterion is derived from the test itself (Longford, Holland, & Thayer, 1993).

Observed variable models for assessing measurement equivalence

The CFA- and IRT- based methods, i.e., the two aforementioned bias detection, are considered "latent variable models" because they are designed to estimate psychological and latent constructs. Both techniques diagnose biased items by matching respondents from different cultural groups on estimates of psychological attributes. They differ in the nature of the relationship between the item scores and the underlying latent constructs. The CFA-based approach estimates linear relationships, whereas the IRT-based method assumes nonlinear relationships.

Several non-latent observed variable models have also been developed. A popular method is the Mantel-Haenszel procedure (Holland & Thayer, 1988), which has been thoroughly studied over the past several decades. It was conceptualized and developed especially for detecting item bias in educational test data, and it mainly focuses on dichotomous responses such as "correct" or "incorrect" and "right" or "wrong." Using the total score as a proxy matching variable, this method finds item bias by identifying whether individuals with the same ability level but different cultural backgrounds have an equal chance of success on an item. Despite its popularity, the Mantel-Haenszel procedure has several limitations: the method can mainly handle only dichotomous response data, and it can only detect uniform bias (Swaminathan & Rogers, 1990).

Another observed variable model has been proposed by van de Vijver and Leung (1997). This model uses an analysis of variance (ANOVA) framework and is therefore applicable to Likert-type item responses. Historically, the ANOVA bias-detection technique was first developed by Cleary and Hilton (1968), but was later found to be flawed (Camilli & Shepard, 1994) because it does not follow the "matching principle," which states that item bias is determined relative to the overall ability or attribute level. Van de Vijver and Leung's ANOVA approach uses the total score as a proxy for the matching variable, which is similar to the

Mantel-Haenszel procedure.

To illustrate this point, consider that the total score is divided into several score groups under the assumption that the score is continuous. A two-way ANOVA with two categorical factors, i.e., a "score group" and a "cultural group," and an interaction term is performed on each item. When the interaction between the "score group" and "cultural group" is statistically significant, an item is considered to have nonuniform bias. A significant main effect of "cultural group" indicates uniform item bias, and a test for the main effect of "score group" is of no particular interest. Van de Vijver and Leung's ANOVA bias-detection model has a number of positive attributes: First, it is based on the matching principle that a biased item is determined in relation to the person's ability level. Second, the approach can detect both uniform and nonuniform item bias. Third, the ANOVA approach can be applied to polytomous Likert-type items.

In this ANOVA model, the formulation of score groups is crucial, and it appears to significantly determine to bias-detection success. As I previously explained, the method divides the total score into several score groups, but the definition of score groups is a rather ambiguous and uncertain process. Based on the premise that refined score groups allow for more powerful statistical analyses, van de Vijver and Leung (1997) make the following recommendations for forming score groups: (a) score groups should have at least 50 subjects and (b) the number of subjects included in groups should be nearly equal. However, these recommendations are not easy to follow in actual studies. The latter recommendation is particularly difficult because subjects' responses for some options are often clustered around the overall mean.

One of the difficulties with score groups stems from categorizing the total score into several score groups. The dichotomizing of a continuous variable reduces the amount of variance that can be accounted for and limits the statistical power of the analysis. Cohen (1983) discusses the cost of dichotomizing variables and suggests that it results in a loss of one-fifth to two-thirds of the variance that may be accounted for by the original variables. Moreover, as a result of dichotomizing normally distributed variables, it induces measurement error and hence, erroneous description of the data “because all the cases coded as being at a single value of the artificial dichotomy actually have substantially different true scores” (Cohen, Cohen, West, & Aikin, 2003, p. 298). The ANOVA model is simple and easy to use, which increases the stability and capability of the bias analysis, but the continuous nature of the total score be retained and used intact. One way to approach this requirement is to model the total score as a continuous variable in the GLM framework, which will be discussed in the following section.

General linear model

GLM is an umbrella term that refers to linear models such as ANOVA, analysis of covariance (ANCOVA), or regression. In a GLM, the dependent variable, which is usually a continuous variable, is linearly combined with more than one independent variable. It is customary to call the analysis an ANOVA if independent variables are categorical and a multiple regression (MR) if the independent variables are continuous. When independent variables are both continuous and categorical, the analysis can be called an ANCOVA, an aptitude treatment interaction (ATI), or a trait treatment interaction (TTI). These models share a common goal, which is to “adjust” or “equate” categorical variable groups with respect to the relevant continuous variables that differ between them (Pedhazur, 1997). In an ANCOVA model, the relevant continuous variable is referred to as a covariate or concomitant variable.

The present study uses a linear model with continuous and categorical variables to test DIF. The model testing bias for item *i* is given by the following equation:

$$O_i = \mu_i + \alpha_i + \beta_i + \alpha_i\beta_i + \varepsilon_i$$

where O_i is the item response, μ_i is the overall mean, α_i is group membership, β_i is the total score, $\alpha_i\beta_i$ is the group membership and the total score interaction term, and ε_i is the error term. The error is normally and independently distributed with a mean of zero and constant variance. Statistical tests are conducted on three regression weights: α_i, β_i, and $\alpha_i\beta_i$. The β_i weight, the psychological trait tapped by the total score, represents the degree to which individuals with higher total scores tend to choose higher response options on a given item. Because it is almost always statistically significant, less interest is given to this weight, as it is the case in van de Vijver and Leung's ANOVA model. When the α_i weight differs significantly from zero, the mean of item response is statistically different between groups after controlling for the effect of β_i in the equation given above. This is the case with uniform bias because at the trait level, the item response of one group is significantly higher or lower than that of the other group. The presence of nonuniform bias is indicated when the interaction term, $\alpha_i\beta_i$, is statistically significant, which is the case when the relationship between the item and the construct varies across groups, thus indicating that the item is more salient for one group than another (Orlando & Marshall, 2002).

In this study, I will test the GLM-based method for bias detection. The linear model that I will test uses the total score as a continuous variable without grouping responses into several score group categories, which is an extension of the perspective of van de Vijver and Leung's ANOVA model (1997). Using simulation data, I examine the GLM model's effectiveness by comparing it to the CFA approach, which is one of the most promising techniques among current, state-of-the-art methodologies.

Method

This study uses simulated data. Monte Carlo simulation studies enable us to "know" parameters in advance so that we can test how effectively detection methods flag biased items. In this study, a 20-item test with 5 category options for the reference and focal groups is simulated with two sample sizes: 500 and 1,000 per group. This 20-item unidimensional test has 4 biased items and 16 unbiased items. The number of biased items is minimized because of the recommendation of Harwell and colleagues (1996) that the ratio of DIF items to unbiased items should be less than one-fifth. The sample size varies (N=500 =and N=1,000) because similar simulation studies (e.g. Stark, Chernyshenko, & Drasgow, 2006) suggest that the sample size impacts the model's ability to detect biased items.

Data properties

The simulation data were obtained using the computer program WINGEN2 (Han & Hambleton, 2007). This program generates data using several IRT models. To obtain polytomous Likert-type data, the graded response model (GRM) was utilized. GRM is a polytomous IRT model that was developed by Samejima (1968).

I undertook the data-simulation process using the following three steps: (1) generating ability values, (2) generating item-parameters values, and (3) generating item-response data. In this study, ability parameters were fixed with a mean of zero and a standard deviation of one for the reference and the focal groups to avoid the risk of confounding the results with theta differences.

To obtain item parameter values, I first generated item parameters for the reference group. Item location parameters (b_{jk}) were randomly sampled from a normal distribution with a mean of zero and a standard deviation of one. Item discrimination parameters (a_j) were randomly sampled from a uniform distribution with a range of 1.0 to 2.0. Location and discrimination parameters are analogous to factor loadings and intercepts in factor analytic models (Ferrando, 1996).

Some item parameters were then changed to create DIF items in the focal group. Via modification of the a_j or b_{jk} values in the reference group, the first four items in the focal group were simulated to show bias. Item 1 and Item 2 are nonuniform DIF items for which only item discrimination parameters were changed. For Item 3 and Item 4, only the location parameters were arranged to show uniform DIF.

In this study, the magnitude of the differences between biased-item values was manipulated to be 0.5 (the low-biased condition) or 1.0 (the high-biased condition). Item 1 and Item 3 were low-magnitude DIF items. Item 1's discrimination parameter and Item 3's 4 location parameters were reduced by 0.5 from the simulated item values in the reference group. Similarly, the values for Item2's discrimination parameter and Item4's 4 location parameters in the focal group, which were high-magnitude biased items, were obtained by subtracting 1.0 from the corresponding item values in the reference group. Aside from these biased items, the rest of the item values were exactly the same in the reference and focal groups (see Table 1).

This study also manipulated the sample size by creating two sample conditions, i.e., a small-sample condition (*N*=500) and a large-sample condition (*N*=1,000), for the reference and focal groups. These sample sizes were selected in accordance with past studies (e.g. Meade & Lautenschlager, 2004). Finally, twenty-five of item-response data were replicated for each sample-size condition based upon these simulated item and ability values.

Data analysis

I examined the proposed GLM approach's ability to detect biased items by comparing it with the CFA method. The GLM approach was analyzed using the GLM function implemented in the SPSS computer program (SPSS, 2006). The dependent variable was the individual item's response, the "factor" was group membership (i.e. the reference or focal groups) and the "covariate" was the total score. A significant interaction between "covariate" and "factor" indicates a nonuniform DIF because the item has higher (or lower) discrimination in one group than in the other. When the item score was not invariant across groups (based on the total score), a uniform DIF was revealed by a significant main effect of "factor." For all significant tests, the Type-I error was controlled by setting alpha=.05.

In observed variable methods, biased items must be filtered out from the internal matching variable, as it is a common practice to "purify" the total score with preliminary item screening in the Mantel-Haenszel procedure (Millsap, 2011). Sixteen unbiased items were used to obtain the total score because the biased items were known in advance. To calculate the score, I summed the item responses from these unbiased items. However, the studied item must be included in the total score regardless of whether it is biased. A past simulation study using the Mantel-

Table 1. Item parameters that were generated and used to simulate the datasets

	Parameters									
	Reference group					Focal group				
Item	a	b1	b2	b3	b4	a	b1	b2	b3	b4
1	1.528	−2.410	−1.032	−0.383	0.350	1.028	−2.410	−1.032	−0.383	0.350
2	1.598	−0.753	−0.171	0.154	0.937	0.598	−0.753	−0.171	0.154	0.937
3	1.609	−0.608	0.027	0.222	1.298	1.609	−1.108	−0.473	−0.278	0.798
4	1.909	−0.271	0.196	0.512	1.442	1.909	−1.271	−0.804	−0.488	0.442
5	1.646	−0.505	−0.491	−0.239	0.172	1.646	−0.505	−0.491	−0.239	0.172
6	1.631	−0.892	−0.185	−0.022	0.760	1.631	−0.892	−0.185	−0.022	0.760
7	1.894	−0.919	0.092	1.060	1.696	1.894	−0.919	0.092	1.060	1.696
8	1.092	−2.140	−1.925	−0.543	1.851	1.092	−2.140	−1.925	−0.543	1.851
9	1.094	−1.624	−0.989	−0.629	−0.064	1.094	−1.624	−0.989	−0.629	−0.064
10	1.135	0.289	0.679	0.687	1.511	1.135	0.289	0.679	0.687	1.511
11	1.081	−1.238	−1.108	−1.105	1.068	1.081	−1.238	−1.108	−1.105	1.068
12	1.687	−1.367	−1.343	−0.822	2.244	1.687	−1.367	−1.343	−0.822	2.244
13	1.426	−0.910	−0.406	0.046	0.880	1.426	−0.910	−0.406	0.046	0.880
14	1.332	−1.073	−0.616	0.343	0.493	1.332	−1.073	−0.616	0.343	0.493
15	1.654	−1.278	−1.166	−0.327	1.071	1.654	−1.278	−1.166	−0.327	1.071
16	1.325	−0.762	−0.741	−0.006	0.738	1.325	−0.762	−0.741	−0.006	0.738
17	1.798	−1.119	−0.052	0.489	1.249	1.798	−1.119	−0.052	0.489	1.249
18	1.672	−1.224	−0.174	−0.100	0.684	1.672	−1.224	−0.174	−0.100	0.684
19	1.337	−1.565	−1.038	0.392	1.060	1.337	−1.565	−1.038	0.392	1.060
20	1.213	−1.239	−0.506	0.059	0.958	1.213	−1.239	−0.506	0.059	0.958

Haenszel procedure suggests that the total score exclusive of the studied item was likely to indicate DIF even though no bias was actually present (Donoghue, Holland, & Thayer, 1993). As a result, when I examined the four biased items, the item responses from the studied and biased items were added to the total score that was calculated from the sixteen unbiased items.

The same data were also analyzed with a confirmatory factor analysis likelihood ratio (CFA-LR) test. For this analysis, I utilized the simultaneous and multigroup function implemented in Amos 18.0J (Arbuckle, 2009). Following the procedures outlined by Oort (1998) and further elaborated by Chan (2000), biased items were examined with a series of likelihood ratio tests. Using chi-square statistics, I compared a restricted model in which I imposed equality constraints on item parameters across group with a baseline model. The baseline model had no constraints on parameters except for the following, which were used for identification purposes: Item 19, an unbiased item, was used as a reference indicator whose factor loading was set to 1, and the intercept was equalized across groups. When testing bias for Item 19, I used another unbiased item (Item 20) as the reference indicator. Similarly to the simulation in the data generation, the mean and standard deviation of the baseline model's latent factor were, respectively, fixed at zero and one for both groups. When a significant difference between the baseline and restricted models' chi-square values is observed, there is invariance where the restricted model's equality constraints are placed. Therefore, uniform bias was indicated when restricted models with equal intercepts significantly departed from their baseline models. Similarly, nonuniform bias was indicated when restricted models with equal factor loadings were significantly different from the baseline model.

Results

Table 2 shows the number of biased and unbiased items detected by the GLM and CFA-LR approaches at alpha=.05. For example, in the *N*=500 condition, the number of uniform DIF items detected was 3 out of 25 replications (12.0%), which was simulated to show a low-magni-

Table 2. The number of DIF items detected and the false positive rate

		Methods							
		GLM				CFA-LR			
Sample size	Bias magnitude	Nonuniform bias	False positive	Uniform bias	False positive	Nonuniform bias	False positive	Uniform bias	False positive
500	0.5	3/25 (12.0%)	7/400 (1.8%)	25/25 (100%)	16/400 (4.0%)	4/25 (16.0%)	10/400 (2.5%)	25/25 (100%)	11/400 (2.8%)
	1.0	24/25 (96.0%)		25/25 (100%)		25/25 (100%)		25/25 (100%)	
1,000	0.5	10/25 (40.0%)	10/400 (2.5%)	25/25 (100%)	17/400 (4.3%)	12/25 (48.0%)	6/400 (1.5%)	25/25 (100%)	5/400 (1.3%)
	1.0	25/25 (100%)		25/25 (100%)		25/25 (100%)		25/25 (100%)	

tude (i.e. 0.5), uniform bias for Item 1. The total number of unbiased items is 400 over 25 replications because each replication has 16 unbiased items (16 items × 25 replications). Again, in the N=500 condition, I falsely detected 7 unbiased items as biased items, which is a 1.8% false-positive rate.

The general findings of the GLM method were as follows: (a) detection rates in the large-sample condition were higher than in the small-sample condition, (b) items with a large amount of bias were flagged more than items with a small amount of bias, (c) nonuniform DIF items were better detected than uniform DIF items, and (d) false-positive rates were less than 5% and about equal across all conditions; however, they tended to appear to be more inflated for nonuniform DIF items, particularly when the sample size was large. I observed results that are similar to (a), (b), (c) and (d) in the CFA-LR approach. Overall, the GLM and CFA-LR methods exhibited remarkably similar bias-detection patterns.

Differences between two methods

When I looked closely at the results, I noticed interesting similarities and differences between the GLM and CFA-LR methods. For uniform DIF items, I found perfect detection using both approaches. They are both successful at detecting uniform DIF (even with the severe and unfavorable conditions of the low-magnitude and/or small-sample conditions).

For the results of nonuniform DIF items, the power was decreased in both approaches, although I observed slight differences across sample sizes and bias magnitudes. In the small-sample size and low-magnitude conditions, the detection rate of the GLM was 12%, which was slightly lower than that of the CFA-LR approach (16%). The detection rate for nonuniform DIF improved with the larger sample size. The GLM detected 10 biased items out of 25 (40%), which again was slightly lower than the CFA-LR for the same condition (48%). The power also improved as the bias grew. For the large amount of nonuniform DIF items, both the GLM and CFA-LR methods exhibited perfect detection except in the small-sample-size condition for the GLM, where the models detected 24 out of 25 nonuniform DIF items (96%).

Overall, the GLM and CFA-LR methods have a fairly similar efficacy for detecting nonuniform bias. The nonuniform DIF detection accuracy seems to be influenced by sample-size or bias-magnitude factors. In contrast to the uniform bias, the nonuniform DIF detection capability deteriorated in both approaches. The power to detect nonuniform DIF was reduced, particularly in the small-sample and low-magnitude bias conditions. Although similar patterns were observed, this trend was more pronounced in the GLM approach than in the CFA-LR approach.

False-positive rates

In this study, the total number of unbiased items tested was 400 as a result of testing 16 unbiased items over 25 replications. The false-positive rates were all less than 5% with a range from 1.3% to 4.3%. Although they were all under the expected value, the type-I error rates for uniform DIF detection were slightly higher than those for nonuniform DIF detection with the GLM method. The GLM the error rates for uniform DIF were also slightly inflated compared with CFA-LR. The sample size has no major impact on error rates, although the false-positive rates of the GLM appear to be slightly increased in the large-sample condition.

Discussion

For cross-cultural researchers, measurement equivalence has been a central concern for many years. Without knowing the psychometric properties of test items, the results of cross-cultural comparisons are ambiguous and perhaps misleading. Although these measurement issues are important, a preliminary item-bias analysis is not yet a prevalent practice in the cross-cultural research community, partly because of inherent difficulties in conducting an item analysis. Some researchers have recommended conducting an item analysis using latent-variable models, such as CFA or IRT (Meade & Lautenschlager, 2004; Stark, Chernyshenko, & Drasgow, 2006). However, these models may not always to be the best choice for applied researchers who face various challenges in actual studies. Some of these challenges include the following: Latent-variable models often require many subjects for stable parameter estimation, and specialized commercial computer software, such as LISREL and MULTILOG, is needed to fit models to the data, and above all, some psychometrics skills and knowledge are required when using these statistically advanced models.

In this study, I used the liner model framework to explore a simpler, more user-friendly alternative for bias detection. Based on simulated data, I examined the accuracy of the proposed GLM method by comparing it to the accuracy of the CFA-LR method. Overall, this study demonstrated the efficacy of the GLM method; its bias-detection capability appears to be satisfactory in most conditions, which was quite similar to the CFA-LR results. Across various conditions, the type-I error rates were all near or below the expected alpha=.05 value. I highlight some of the characteristics of the GLM procedure below.

I have demonstrated that the GLM and the CFA-LR both perfectly detected uniform item bias. Their detection capability for uniform DIF seems not to be influenced by a sample-size or bias-magnitude decrease. Therefore, GLM, similarly to CFA-LR, detected uniform bias fairly well. Both approaches also demonstrated similar patterns for nonuniform bias detection. Overall, both approaches performed very well with high-magnitude and large-sample conditions. However, a reduction of the sample size and bias magnitude leads to decreased power for detecting nonuniform DIF using both methods, and the trend appears to be more pronounced for GLM than for CFA-LR.

Although the nonuniform-bias-detection capability using the GLM and CFA-LR is rather poor under some conditions, these results are somewhat expected considering similar past simulation studies. For example, in Wanichtanom’s Monte Carlo study (2001), where the area procedure (Raju, 1988) was used to create bias items, the overall detection rate of CFA-LR for nonuniform bias was 56%. For the sample-size condition where N=1,000, the detection rates differed for nonuniform-bias magnitudes that were low (approximately 0.12), medium (approximately 0.38), and high (approximately 1.10) (36%, 64%, and 69%, respectively). In another simulation study, this time for the logistic regression procedure (Swaminathan & Rogers, 1990), approximately 50% of nonuniform bias items were detected in a small-sample (*N*=250), short-test (40 items) condition, and approximately 75% were detected in a large-sample (*N*=500), long-test condition (80 items). This study's significance level was set at alpha=.01.

The statistical power for detecting interaction effects has long been a concern in social and behavioral sciences. In the ANOVA or regression analysis, the detection of interactions is much more difficult than the detection of main effects (Cronbach & Snow, 1977; Cohen, Cohen, West, & Aikin, 2003). As with the detection of nonuniform bias, the nature of the interaction seems to impact the statistical power. For example, I mentioned that the overall detection rate for nonuniform bias was 56% in Wanichtanom's simulation work (2001). However, consideration of the nature of the interaction reveals a completely different picture, as the detection rates for ordinal nonuniform bias were much higher the detection rates for disordinal nonuniform bias. For the medium- to high- nonuniform-bias magnitudes, perfect or nearly perfect detection was observed. Even for low-magnitude nonuniform items, 36% to 64% of these ordinal biased-items were detected. On the other hand, the detection rates for disordinal nonuniform bias were quite low. Over 25 data replications and 3 bias-magnitude conditions, 2 items (8%) in the low-, 0 item (0%) in the medium-, and 2 items (8%) in the high-magnitude conditions were detected. Lower power for nonuniform bias detection was also indicated in another logistic regression procedure simulation study. DIF items with medium difficulty where item characteristic curves from two groups that intersected in the middle range of the ability parameter were difficult to detect (Rogers & Swaminathan, 1993). These results indicate that the nature of the interaction impacts the ability to detect nonuniform bias. Further simulation work will be needed to identify how the form of bias affects the statistical power for detecting nonuniform bias with the GLM method.

Advantages of the GLM method

The GLM approach has a number of positive features in addition to basic functional advantages, such as the identification of both uniform and nonuniform types of bias. First, the GLM approach is flexible in its modeling; it assumes that the total score is an internal criterion for flagging bias. Although the internal criterion method is popular, it is also potentially circular (Camilli & Shepard, 1994). Furthermore, the use of an internal criterion is, by definition, "self-norming" (Longford, Holland, & Thayer, 1989), which allows us to identify only relative discrepancies in item properties across groups because there is no absolute way to filter out biases from other test items. To avoid these potential problems of the internal criterion method, the incorporation of an external criterion is necessary. The GLM modeling method is flexible and can be easily expanded to include other external ability criteria.

Additionally, as a linear-model application, the GLM approach offers effect-size estimation. The provision of effect-size measures is useful for evaluating the "practical significance" of biased items. An interpretable measure of the bias magnitude is also needed in DIF research because the sensitivity of a statistical hypothesis test depends on sample size (Potenza & Dorans, 1995). In GLM bias detection, bias magnitudes can be evaluated through R^2 difference tests. Starting with the baseline model, where only the matching variable (i.e. the total score) is modeled, we can examine incremental amounts of variance that are explained by the membership variable and interaction term. As has been shown in studies on the role of effect size measures in the logistic regression DIF procedure (e.g. Jordoin & Gierl, 2001), the availability of effect size measures greatly enhances the practicality of the GLM procedure.

Another useful feature of the GLM method is the ability to search for anchor items prior to the implementation of a more sophisticated bias analysis, such as CFA-LR or IRT. The LR test approach is more powerful than the bias indices approach (Orlando & Marshall, 2002), but success with the LR test approach depends on the establishment of stable anchor items (Millsap, 2011). Several iterative procedures have been proposed to search for anchor items for multigroup CFA (e.g. Rensvold

& Cheung, 2001). However, these procedures require multiple iterations of significance testing and thereby carry a heightened risk for type-I errors. The iteration process requires several model comparisons, and there is some uncertainty regarding the rules for accepting or rejecting models (Millsap, 2011). The GLM method is a simple but powerful alternative that does not entail laborious, iterative processes to search for anchor variables. In fact, prior to conducting IRT-based DIF detection, Orlando and Marshall (2002) applied logistic regression procedures to find anchor items, which worked fairly well in detecting DIF items in the subsequent IRT analyses. As I previously mentioned, the current GLM-based bias analysis is theoretically and methodologically equivalent, except for the type of the dependent variables used in the logistic regression approach.

Recommendations

Before concluding, I offer some caveats from a practical viewpoint about the total score because it appears that the nature of the total score significantly impacts bias-detection capability. The GLM approach uses the total score as a proxy for matching subjects who have different cultural backgrounds. This internal matching variable must be "purified" prior to use because the use of a total score that includes biased items may lead to erroneous results (Doran & Holland, 1993; Holland & Thayer, 1988). To filter out biased items, Holland and Thayer (1988), for example, recommended a two-stage procedure for the Mantel-Haenszel approach. In this refinement process, each item is first studied using the total score across all items. Then the total score is re-calculated without the biased items that were detected during the first stage. This refined total score will be used as the internal criterion for the bias detection that follows. Note that the studied item must be included in the total score (Donoghue, Holland, & Thayer, 1993), regardless of whether the item is flagged as biased in the first stage of the purification process. Therefore, depending on which item is examined, the total score may need to be redefined (Millsap, 2011).

Another concern about the total score is its dimensionality because “the unidimensionality of the matching variable is central to the DIF assessment process” (Doran & Holland, 1993, p. 61). The bias-detection technique for observed variable methods assumes that the proxy criterion variable reflects the underlying ability, or aptitude, that the whole test intends to measure. For a unidimensional test, which is implicitly and explicitly assumed for many statistical bias-detection methods (Camili & Shepard, 1994), secondary factors or irrelevant variables should be excluded from the total score; otherwise, the bias-detection results will be misleading. Consequently, this dimensionality assumption has to be confirmed in the GLM method with a preliminary analysis that uses factor analytic techniques (e.g. exploratory factor analysis).

In conclusion, this simulation study identifies the effectiveness of the GLM method for bias detection and suggests that the bias-detection capability of the GLM method is fairly similar to that of the CFA-LR. The GLM will be a valid alternative, particularly for the case where advanced statistical methods, such as CFA-LR and IRT, are not available, but further research is also needed to identify how other variables such as test length and biased- to nonbiased- item ratio, affect power and type-I errors. I hope that this simple, practical bias-detection technique allows more researchers to easily examine measurement equivalence and enhances the validity of cross-cultural investigations.

References

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In: P.W. Holland & H.Wainer, eds. *Differential Item Functioning*, pp. 3–23. Hillsdale, NJ: Erlbaum.

Arbuckle, J. L. (2009). *AMOS 18 User’s Guide*. Chicago: Amos Development Corporation.

Byrne, B. M., & Watkins, D. (2003). The issue of measurement invariance revised. *Journal of Cross-Cultural Psychology*, *34*, 155–175.

Camilli, G. & Shepard, L. A. (1994). *Methods for Identifying Biased Test Items*. Thousand Orks, CA: Sage.

Chan, D. (2000). Detection of differential item functioning on the Kirton Adaption-Innovation Inventory using multiple-group mean and covariance structure analyses. *Multivariate Behavioral Research*, *35*, 169–199.

Cleary, T. A. & Hilton, T. L. (1968). An investigation of item bias. *Educational and Psychological Measurement*, *28*, 61–75.

Cohen (1983). The cost of dichotomization. *Applied Psychological Measurement*, 7, 249–253.

Cohen, J., Cohen, P., West, S. G., & Aikin, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Mahwah, NJ: Lawrence Erlbaum.

Cronbach, L. J. & Snow, R. E. (1977). *Aptitudes and Instructional Methods*. New York: John Wiley & Sons.

Donoghue, J. R., Holland, P. W., & Thayer, D. T. (1993). A Monte Carlo study of factors that affect the Mentel-Haenszel and standardization measures of differential item functioning. In: H. Wainer & H. I. Braun, eds. *Test Validity,* pp. 137–166. Hillsdale, NJ: Lawrence Erlbaum.

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

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. *Psychological Bulletin*, *95*, 134–135.

Ferrando, P. J. (1996). Calibration of invariant item parameters in a continuous item response model using the extended Lisrel measurement submodels. *Multivariate Behavioral Research*, *31*, 419–439.

Gudykunst, W. B., Matsumoto, Y., Ting-Toomy, S., Nishida, T., Kim, K., & Heyman, S. (1996). The influence of cultural individualism-collectivism, self-construals, and individual values on communication styles across cultures. *Human Communication Research*, *22*, 510–543.

Häder, S., & Gabler, S. (2003). Sampling and estimation. In: J. A. Harkness, P. P. Mohler & F. J. R. van de Vijver, eEds. *Cross-cultural Survey Methods,* pp. 117–134. Hoboken, NJ: Wily-Interscience.

Han, K. T., & Hambleton, R. K. (2007). *User's Manual for WinGen*. http://www.umass.edu/remp/software/simcata/wingen/WinGen_Manual_Han_Hambleton_2007.pdf

Harwell, M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte carlo studies in item response theory. *Applied Psychological Measurement*, *20*, 101–125.

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. Hillsdale, NJ: Lawrence Erlbaum.

Jordoin, M. G. & Gierl, M. J. (2001). Evaluating power and type I error rates using an effect size with the logistic regression procedure for DIF. *Applied Measurement in Education*, *14*, 329–439.

Jöreskog, K. G. & Sörbom, D. (2002). *LISREL 8.52 for Windows*. [computer manual]. Lincolnwood, IL: Scientific Software.

Kankaraš, M. & Moors, G. (2011). Measurement equivalence and extreme response bias in the comparison of attitudes across Europe. *Methodology*, 7, 68–80.

Kleinknecht, R. A., Dinnel, D. L., Kleinknecht, E. E., Hiruma, N., & Harada, N. (1997). Cultural factors in social anxiety: A comparison of social phobia symptoms and taijin kyofusho. *Journal of Anxiety Disorders*, *11*, 157–177.

Levine, T. R., Bresnahan, M. J., Park, H. S., Lapinski, M. K., Wittenbaum, G. M., Shearman, S. M., Lee, S. Y., Chung, D., & Ohashi, R. (2003). Self-construal scales lack validity. *Human Communication Research*, *29*, 210–252.

Longford, N. T., Holland, P. W., & Thayer, D. T. (1993). Stability of the MH D-DIF statistics across populations. In: P. W. Holland & H. Wainer, eds. *Differential Item Functioning,* pp. 171–196. Hillsdale, NJ: Erlbaum.

McClelland, G. H. & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin*, *114*, 376–390.

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, *98*, 224–253.

Matsumoto, D. (1999). Culture and self: An empirical assessment of Markus and Kitayama's theory of independent and interdependent self-construals. *Asian Journal of Social Psychology*, *2*, 289–310.

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, *7*, 361–388.

Millsap, R. E. (2011). *Statistical Approaches to Measurement Invariance*. NY: Taylor & Francis.

Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling*, *5*, 107–124.

Orlando, M., & Marshall, G. N. (2002). Differential item functioning in a Spanish translation of the PTSD checklist: Detection and evaluation of impact. *Psychological Assessment*, *14*, 50–59.

Pedhazur, E. J. (1997). *Multiple Regression in Behavioral Research*. London: Thomson Learning.

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

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement,* No. 17.

SPSS (2006). SPSS for Windows (Version 14.0J). [computer program]. Tokyo: SPSS Japan Inc.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting DIF with CFA and IRT: Toward a unified strategy. *Journal of Applied Psychology*, *91*, 1292–1306.

Steenkamp, J. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*, 78–90.

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

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *56*, 365–379.

Rogers, H. J. & Swaminathan, H. (1993). A comparison of the logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, *17*, 105–116.

Rensvold, R. B. & Cheung, G. W. (2001). Testing for metric invariance using structural equation models: Solving the standardization problem. In: C. A. Schriesheim & L. L. Neider, eds. *Equivalence in Management*, vol. 1, pp. 25–50. Greenwich, CN: Information Age Publishing.

Thissen, D. (1991). *MULTILOG User's Guide* (*Version 6.0*) [computer manual]. Mooresville, IL: Scientific Software.

Van de Vijver, F. L. R., & Leung, K. (1997). *Methods and Data Analysis for Cross-Cultural Research*. Thousand Oaks, CA: Sage.

Wanichtanom, R. (2001). *Methods of Detecting Differential Item Functioning: A Comparison of Item Response Theory and Confirmatory Factor Analysis*. Unpublished dissertation, Old Dominion University.

Welkenhuysen-Gybels, J., Billet, J., & Cambré, B. (2003). Adjustment for acquiescence in the assessment of the construct equivalence of Likert-type score items. *Journal of Cross-Cultural Psychology*, *34*, 702–722.

Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning* (*DIF*)*: Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type* (*Ordinal*) *Item Scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.