Examining Measurement Invariance and Differential Item Functioning With Discrete Latent Construct Indicators: A Note on a Multiple Testing Procedure

Educational and Psychological
Measurement
2018, Vol. 78(2) 343–352
© The Author(s) 2016
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0013164416670984
journals.sagepub.com/home/epm

Tenko Raykov¹, Dimiter M. Dimitrov^{2,3},
George A. Marcoulides⁴, Tatyana Li¹, and
Natalja Menold⁵

Abstract

A latent variable modeling method for studying measurement invariance when evaluating latent constructs with multiple binary or binary scored items with no guessing is outlined. The approach extends the continuous indicator procedure described by Raykov and colleagues, utilizes similarly the false discovery rate approach to multiple testing, and permits one to locate violations of measurement invariance in loading or threshold parameters. The discussed method does not require selection of a reference observed variable and is directly applicable for studying differential item functioning with one- or two-parameter item response models. The extended procedure is illustrated on an empirical data set.

Keywords

binary item, differential item functioning, false discovery rate, latent variable modeling, measurement invariance, model identification, multiple testing

¹Michigan State University, East Lansing, MI, USA

²George Mason University, Fairfax, VA, USA

³National Center for Assessment, Riyadh, Saudi Arabia

⁴University of California, Santa Barbara, CA, USA

⁵Leibniz Institute for the Social Sciences, Mannheim, Germany

Corresponding Author:

Tenko Raykov, Measurement and Quantitative Methods, Michigan State University, 443A Erickson Hall, East Lansing, MI 48824, USA.

Email: raykov@msu.edu

Measurement invariance (MI) with respect to latent constructs evaluated by multiple indicators has attracted a great deal of attention by quantitative and substantive behavioral and social scientists over the past several decades (e.g., Millsap, 2011). An impressive body of literature on studying MI has accumulated in the past 25 years or so, which documents methodological advances in this area (see also Cheung & Rensvold, 2002; Meredith, 1993; Raykov, Marcoulides, & Li, 2012; Vandenberg & Lance, 2000).

Recently, Raykov, Marcoulides, and Millsap (2013) proposed a method for studying MI, which was developed within the framework of latent variable modeling (Muthén, 2002) and based on the Benjamini–Hochberg (BH) multiple testing procedure (Benjamini & Hochberg, 1995; see, e.g., Raykov, Marcoulides, Lee, & Chang, 2013, for a nontechnical description). A limitation of that method is, however, the fact that it is strictly applicable only with (approximately) continuous manifest measures. Yet latent construct indicators that are highly discrete, such as binary or binary scored items, are very often of concern in behavioral and social research (e.g., de Ayala, 2009). In such settings, the Raykov, Marcoulides, and Millsap (2013) MI testing approach cannot be recommended, as it can potentially yield misleading results and interpretations.

The purpose of this note is to extend the Raykov, Marcoulides, and Millsap (2013) method for examining MI to the case of binary factor indicators (items) with no guessing. The remainder of this discussion makes similarly use of the false discovery rate concept, after accounting for the binary nature of the manifest indicators by fitting pertinent models using an estimation method appropriate for their discrete distribution. This extension broadens substantially the scope of applicability of the Raykov, Marcoulides, and Millsap (2013) MI examination approach, by including also the case of binary/binary scored measures or items that are frequently employed in contemporary educational and psychological studies. Additionally, studying differential item functioning (DIF) is also possible in this setting with the outlined extended MI examination procedure, under a two-parameter item response model or a special case of it (i.e., a one-parameter model; see also Raykov & Marcoulides, 2016).

Background, Notation, and Assumptions

The present article assumes that a set of observed binary or binary scored items with no guessing are given, denoted $\underline{y} = (y_1, y_2, \dots, y_k)'$, which represent the components of a psychometric scale, test, survey, questionnaire, inventory, self-report, subscale, or testlet ($k > 2$, conditional on overall identification of any model version utilized below; underlining denotes vector and priming transposition in this article). We also assume that the set $\underline{y}$ fulfils the configural invariance condition with regard to G ($G > 1$) distinct populations of interest consisting of independent subjects, with large samples being made available from each of them (Millsap, 2011); that is, in each of the G groups (populations), the same number q ($q \geq 1$) of latent factors are being evaluated by (load on) the same measures in $\underline{y}$.

In this setting, a factor analysis model can thus be advanced within each group, whose parameters are (a) the loadings of the assumed underlying normal variables $\underline{y}^*$ on their common factors, collected correspondingly in the $k \times q$ matrices $\Lambda_1, \Lambda_2, \dots, \Lambda_G$ in the groups; and (b) the associated thresholds $\tau_{g,j}$, assembled respectively in the $k \times 1$ vectors $\underline{\tau}_1, \underline{\tau}_2, \dots, \underline{\tau}_G$ in the groups ($j = 1, \dots, k$; cf. Muthén, 1984). Hence, this model in the g th group is

$$\underline{y}_g^* = \Lambda_g \eta_g + \delta_g, \quad (1)$$

where $\underline{y}_g^*$ is the $k \times 1$ vector of underlying normal variables associated with the manifest indicators in that group, η_g is the $q \times 1$ vector of their common factors, and δ_g is the associated $q \times 1$ vector of pertinent unique factors ($g = 1, \dots, G$; e.g., Muthén & Muthén, 2016).

As elaborated in the MI literature, the following couple of multiparameter constraints represent a necessary condition for MI (e.g., Millsap, 2011):

$$\Lambda_1 = \Lambda_2 = \dots = \Lambda_G \quad (2)$$

and

$$\underline{\tau}_1 = \underline{\tau}_2 = \dots = \underline{\tau}_G \quad (3)$$

The pair of Equations (2) and (3) represents also a sufficient condition for lack of DIF under a two-parameter item response theory (IRT) model or a special case of it (such as a one-parameter IRT model; e.g., Asparouhov & Muthén, 2015; Lord, 1980).

A Latent Variable Modeling Procedure for Examining Measurement Invariance and Differential Item Functioning With Binary Items

In a setting related to the one presently considered, Raykov, Marcoulides, and Millsap (2013) outlined a multiple testing-based procedure for studying MI, which is applicable with (approximately) continuous factor indicators. Main desirable features of that method are that (a) it proceeds without the need to select a reference measure from a given set of k observed variables in any of the G groups, (b) accounts for the multiple null hypotheses testing involved, and (c) allows location of possible violations of MI.

Procedure Extension for Binary or Binary Scored Items

An extension of the Raykov, Marcoulides, and Millsap (2013) procedure to the binary or binary scored indicator/item case without guessing is readily obtained as follows in parallel to that earlier method (cf. Millsap, 2011). As in that procedure, in

the first step here one fits the G -group model to the set of k binary items (accounting for their categorical nature, using maximum likelihood with categorical measures; Muthén & Muthén, 2016). In this multigroup model, (a) all loadings and thresholds are correspondingly fixed for group identity; (b) the means and variances of the latent variables are set at 0 and 1, respectively, in the first group only but free in all remaining groups; and (c) if applicable, the off-diagonal elements of the latent covariance matrix are free in all groups (when $q > 1$). That is, in Step 1 Equations (2) and (3) are imposed, the means and variances of all factors are correspondingly set at 0 and 1 only in the first group but left unconstrained in all other groups, and in case $q > 1$ the latent covariance matrix is free in all groups except its main diagonal elements that are fixed at 1 in the first group only (e.g., the reference group). Denote this restricted G -group model version as M_0 .

In the second step, individual thresholds as well as individual loadings are successively freed as single parameters from their group equality in M_0 , leading to $2k$ respective model versions $M(1), \dots, M(2k)$ (in case $q = 1$, and correspondingly with $q > 1$; the rest of this subsection describes in detail the procedure for $q = 1$, with obvious extension for $q > 1$ where one may consider up to $2kq$ such consecutive single-parameter relaxed models; cf. Raykov, Marcoulides, & Millsap, 2013). These $2k$ models have obviously the property that M_0 is nested in each one of them; in fact, M_0 results from any one of them by setting for group identity (in addition) the only parameter that is free from the cross-group equality constraint in $M(1), \dots$, and $M(2k)$, respectively. Therefore, by using the likelihood ratio test (e.g., Bollen, 1989) for the comparison of M_0 with each of the latter $2k$ models, one obtains the respective p value associated with the single-degree-of-freedom test of the pertinent null hypothesis that the corresponding (free) loading or threshold is in fact identical in the groups. This leads in the end to a set of $2k$ in total p values, which result from these $2k$ comparisons of M_0 to each of the more relaxed models $M(1), \dots, M(2k)$ that have 1 less degree of freedom than M_0 .

In the third step, the BH procedure is applied on this set of p values. This is done in order to determine which of the tested $2k$ null hypotheses are to be rejected, if any, which stipulate group identity also in the respective (free) loading or threshold in $M(1), \dots, M(2k)$. (An R-function accomplishing this application of the BH procedure can be found in Raykov, Marcoulides, Lee, et al., 2013.) Accordingly, denote the number of hypotheses to be rejected by r ($r \geq 0$; cf. Wasserman, 2004). If $r = 0$ of them are to be rejected, the necessary condition for MI is declared consistent with the data, and it is suggested that the analyzed data set does not contain sufficient evidence warranting rejection of MI (or DIF, as discussed in the next subsection of this note). Alternatively, if $r > 0$ of these hypotheses are to be rejected, MI does not hold, since its necessary condition(s) (2) or (3) is violated then. In this case, the present method also indicates which particular of the $2k$ null hypothesis is to be rejected, and thus locates the parameter and item that is not group invariant (i.e., exhibits DIF; see next subsection).

Studying Differential Item Functioning With Two-Parameter or One-Parameter Item Response Models

As an implication of the DIF related discussions, for instance, in Lord (1980), the restrictions presented in the above pair of Equations (2) and (3) are readily found to represent, as a couple, a sufficient condition for lack of DIF in case a two-parameter IRT model is correct for a given set of binary items with no guessing (or a one-parameter IRT model is correct; see, e.g., also Raykov & Marcoulides, 2016, on the equivalence and close relationships of two-parameter IRT models to single-factor models in the setting under consideration in this note, as well as corresponding parameter conversion formulas). Therefore, the method outlined in the preceding subsection of this note is directly applicable for studying DIF under a one- or two-parameter IRT model.^{1,2}

An application of the described extended procedure for MI—and DIF—examination is demonstrated next.

Illustration on Data

For the aims of this section, we use adapted data from a mathematics ability test consisting of $k = 9$ binary items administered to 771 boys and 744 girls (cf. de Boeck & Wilson, 2004). We commence by fitting the single-factor model in each group (see first Mplus command file in the appendix), using the weighted least squares estimation method and accounting for the categorical nature of the nine items (Muthén & Muthén, 2016). This model of unidimensionality is found to be associated with tenable goodness of fit indexes in each of the groups. Specifically, in the reference group (males) these were as follows: $\chi^2 = 58.979$, degrees of freedom (df) = 27, root mean square error of approximation (RMSEA) = .039 with a 90% confidence interval being [.026, .053]; similarly, in the focal group (females), $\chi^2 = 49.086$, $df = 27$, RMSEA = .039 [.018, .048], leading to the suggestion that the administered overall test (nine-item instrument) is homogeneous in both groups.

Next, we fit the two-group model M0, described in detail earlier in this note, using the ML estimation method, and find that it is associated with the following tenable fit indexes as well (see second Mplus command file and notes following it in the appendix): likelihood ratio $\chi^2 = 969.495$, $df = 1,002$, $p = .764$. We then proceed with fitting each of the $2 \times 9 = 18$ models M(1) through M(18) here, with M0 being nested in each one of them, and evaluating the p values pertaining to the test of the null hypothesis that the only loading or threshold parameter free in any of these 18 models is actually also group invariant.³ The results of these analyses are presented in Table 1, with its last column containing the p values of relevance in the next step of the procedure of this article.

Employing the BH multiple testing procedure on these 18 p values, presented in Table 1, leads to the decision to reject one of the pertinent 18 null hypotheses tested, viz. that with p value not exceeding .00006 (for an R-function for BH testing, see, e.g., Raykov, Marcoulides, Lee, et al., 2013). This is the hypothesis associated with the first item of the instrument under consideration. Hence, it is suggested that MI does not hold

Table 1. Likelihood Ratio Test Statistics (χ^2 Values) for the 18 Relaxed Models Used—With Model M0 Nested in Each of Them—and Associated p Values.

Model	Par.	χ^2	df	$\Delta(\chi^2)$	Δdf	p
M(1)	τ_1	953.337	1,001	16.158	1	.000
M(2)	τ_2	968.482	1,001	1.013	1	.031
M(3)	τ_3	964.633	1,001	4.862	1	.027
M(4)	τ_4	965.039	1,001	4.456	1	.035
M(5)	τ_5	964.406	1,001	5.089	1	.024
M(6)	τ_6	969.117	1,001	2.378	1	.012
M(7)	τ_7	965.516	1,001	3.979	1	.046
M(8)	τ_8	966.447	1,001	3.048	1	.081
M(9)	τ_9	966.108	1,001	3.387	1	.066
M(10)	λ_1	968.417	1,001	1.078	1	.299
M(11)	λ_2	967.925	1,001	1.570	1	.210
M(12)	λ_3	969.396	1,001	0.099	1	.753
M(13)	λ_4	960.815	1,001	8.680	1	.003
M(14)	λ_5	962.169	1,001	7.326	1	.007
M(15)	λ_6	969.352	1,001	0.143	1	.701
M(16)	λ_7	969.481	1,001	0.014	1	.906
M(17)	λ_8	965.063	1,001	4.432	1	.035
M(18)	λ_9	969.178	1,001	0.317	1	.573

Note. Par. = threshold or loading released singly in pertinent model (in same row), in which M0 is nested; χ^2 = likelihood ratio test's chi-square value for fitted relaxed model (with corresponding single loading or threshold parameter free, relative to M0); df = degrees of freedom; $\Delta(\chi^2)$ = chi-square difference test statistic, relative to M0 (with chi-square = 969.495, df = 1002; see main text); Δdf = difference in degrees of freedom (relative to M0); p = p value associated with chi-square difference for pertinent model (fitted relaxed version of M0); see main text for goodness-of-fit statistics associated with model M0.

across gender for the used mathematics ability test. Thereby, it is worth stressing that the procedure of this note was able to locate the MI violation in this particular item (Item 1). Applying to its data the popular Mantel–Haenszel test (e.g., Agresti, 2002), we obtain an associated odds ratio of 1.6296, with a 95% confidence interval [1.2594, 2.1085]. Since this odds ratio is markedly (and significantly) above 1, it is suggested that Item 1 shows DIF and thereby favors the focal group, that is, the female group.

Conclusion

Over the past quarter of a century or so, the topic of MI has received an extraordinary amount of attention in the behavioral and social sciences, and in particular in educational and psychological research. An impressively voluminous body of literature has developed during this time in these and cognate disciplines ranging from marketing to sociology. A recently proposed method by Raykov, Marcoulides, and Millsap (2013) offered a means of examining MI without the need for choosing a reference variable, a potentially difficult issue to resolve before commencing MI examination (e.g., Raykov et al., 2012). However, that method was applicable only to multicomponent measuring instruments consisting of (approximately) continuous components.

This note extended the Raykov, Marcoulides, and Millsap (2013) method to the case of binary or binary scored items with no guessing, thus substantially broadening the applicability of that multiple testing-based approach for studying MI. The extended procedure, like that earlier one, permits in addition location of the violation(s) of MI (for a given data set and model). Furthermore, this procedure is directly applicable for examining DIF under one- or two-parameter item response models, and is similarly able to locate items contributing to DIF of an instrument under consideration.

The procedure outlined in this article has several limitations worthwhile pointing out here. First, it is ideally applied when the most restrictive model M_0 is plausible in an empirical setting. Future research is encouraged that would examine what type of misspecifications of this model may still be “tolerated” (and under which conditions) by the present MI examination method, since what may be arguably more relevant then may be instead the plausibility of each of the more relaxed $2k$ models $M(1), \dots, M(2k)$ that model M_0 is nested in. Further, the discussed method in this note, as described above, is applicable in case of no nesting effect in each (sub)population studied. Finally, the procedure is based on asymptotic statistical theory and therefore is best used with large samples. With this in mind, future research is also needed in order to inform when the underlying large-sample theory may obtain practical relevance, with the hope of possibly developing also “rules of thumb” for determining in empirical research sufficient sample sizes in this respect.

In conclusion, the present article offers to educational, behavioral, and social scientists a widely applicable means for studying MI and DIF in multi-item measuring instruments with binary or binary scored items and no guessing. In conjunction with the procedure in Raykov, Marcoulides, and Millsap (2013), the present extension adds to the methodological arsenal of scholars in these and cognate disciplines concerned with binary-, binary scored-, or (approximately) continuous component-based instruments to be examined for MI, or alternatively for DIF under one- or two-parameter IRT models.

Appendix

Mplus Source Codes for Examining Measurement Invariance and Differential Item Functioning With Binary Items (Models $M_0, M_1, \dots, M_{18}$)

```
TITLE:          TESTING SINGLE-FACTOR MODEL PER GROUP.
                MALES GROUP.
DATA:          FILE = MA.DAT;
VARIABLE:      NAMES = ITEM1-ITEM9 FEMALE;
                USEVARIABLES = ITEM1-ITEM9;
                CATEGORICAL = ITEM1-ITEM9;
                USEOBSERVATIONS = FEMALE==0; ! select males group.
MODEL:        MA BY ITEM1-ITEM9; ! MA = math ability.
```

Note. Select females for the subsequent group model testing by using the command “USEOBSERVATIONS = FEMALE ==1;” (or correspondingly modified in other data files with different group naming).

```
TITLE:          FITTING MODEL  $M_0$ .
```

```

DATA:          FILE = MA.DAT;
VARIABLE:      NAMES = ITEM1-ITEM9 FEMALE;
                CATEGORICAL = ITEM1-ITEM9;
                KNOWNCLASS = C (FEMALE = 0 FEMALE = 1) ;
                CLASSES = C (2) ;
ANALYSIS:      ESTIMATOR = ML;
                TYPE = MIXTURE;
                ALGORITHM = INTEGRATION;

MODEL:

%OVERALL%
MA BY ITEM1* (L1)
    ITEM2-ITEM9 (L2-L9) ;
    [ITEM1$1-ITEM9$1] (T1-T9) ;
    [MA@0] ;
MA@1;
%C#2%
MA BY ITEM1* (L1)
    ITEM2-ITEM9 (L2-L9) ;
    [ITEM1$1-ITEM9$1] (T1-T9) ;
    [MA*] ;
MA*;

```

Note. To fit any of the 18 models M(1) through M(18), with M0 being nested in each of them, proceed as follows. For a threshold release, exchange the final 6 lines of the last Mplus command file (for model M0) with the following 6 lines (say for releasing the first threshold):

```

%C#2%
MA BY ITEM1* (L1)
    ITEM2-ITEM9 (L2-L9) ;
    [ITEM2$1-ITEM9$1] (T2-T9) ;
    [ITEM1$1*] ;
    [MA*] ; MA*;

```

For a loading release, exchange the final 6 lines of the last stated full Mplus file (for M0) instead with the following 6 lines (say for releasing the first loading):

```

%C#2%
MA BY ITEM1*
    ITEM2-ITEM9 (L2-L9) ;
    [ITEM1$1-ITEM9$1] (T1-T9) ;
    [MA*] ;
MA*;

```

Authors' Note

This research was in part conducted while Tenko Raykov was visiting the Leibniz Institute for the Social Sciences, Mannheim, Germany.

Acknowledgments

We are grateful to L. Cai and R. J. Wirth for informative discussions on measurement invariance testing, and to C. Wolf and B. Rammstedt for their valuable support.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. The procedure outlined in this article, when used for DIF examination, differs from the DIF study approach in Woods, Cai, and Wang (2013)—as implemented in the software flexMIRT (Houts & Cai, 2015)—in the following two features. In the present procedure, like that in Raykov, Marcoulides, and Millsap (2013), (a) the baseline model is M_0 , which is nested in all key models $M(1)$ through $M(2k)$; and (b) no constraints are utilized at any point that set one or more parameters equal to their estimates obtained with the used data in model M_0 (or in another model). In contrast, the Woods et al. (2013) approach, as implemented in flexMIRT, fixes in the main two-group model, which is distinct from M_0 (and denoted say M^*), the mean and variance in the focal group to their estimates obtained with model M_0 from the same analyzed data set. That approach can then be seen as proceeding by adding successively single slope or intercept group equality constraints in the two-group model M^* , with no further restrictions within or across groups (except the latent mean and variances fixed at 0 and 1 in the reference group, respectively, and in the focal group as mentioned at their estimates from model M_0). With these properties, it may be conjectured that the Woods et al. (2013) approach can be more susceptible to misspecifications in M_0 (particularly when there is DIF in a given data set or study to begin with), which misspecifications seem to be affecting to a lesser degree the procedure outlined in this article; the reason is that (a) no parameter fixing at sample estimates are used here unlike in that alternative, earlier approach to DIF examination and (b) whether model M_0 is tenable for a given data set seems to be empirically less relevant here than in that DIF study approach by Woods et al. (2013; see also Houts & Cai, 2015).
2. The method outlined in this article differs from the DIF evaluation approach in Dimitrov (2016) in that the latter is concerned with quantification of possible aspects of DIF, while the present one (a) is focused on testing for violations of MI and DIF and (b) accounts for multiple testing.
3. The p values in Table 1 are readily obtained using for instance the R-command “1-pchisq (difference in chi-squares, 1),” where “difference in chi-squares” is the discrepancy of the two chi-square values within pertinent row.

References

Agresti, A. (2002). *Categorical data analysis*. New York, NY: Wiley.

- Asparouhov, T., & Muthén, B. (2015). IRT in Mplus (Webnote). Retrieved from <http://www.statmodel.com/download/MplusIRT.pdf>
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57, 289-300.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9, 233-255.
- de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford.
- de Boeck, P., & Wilson, M. (2004). *Explanatory item response models*. New York, NY: Springer.
- Dimitrov, D. M. (2016). *Examining differential functioning of binary items: IRT-based detection in the framework of confirmatory factor analysis* (Technical Report 2016-1). Riyadh, Saudi Arabia: National Center for Assessment.
- Houts, C. R., & Cai, L. (2015). *flexMIRT: Flexible multilevel multidimensional item analysis and test scoring*. Chapel Hill, NC: Vector Psychometric Group.
- Lord, F. M. (1980). *Applications of item response theory to practical settings*. Hillsdale, NJ: Erlbaum.
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58, 525-543.
- Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Taylor & Francis.
- Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49, 115-132.
- Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika*, 29, 81-117.
- Muthén, L. K., & Muthén, B. O. (2016). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.
- Raykov, T., & Marcoulides, G. A. (2016). On the relationship between classical test theory and item response theory: From one to the other and back. *Educational and Psychological Measurement*, 76, 325-338.
- Raykov, T., Marcoulides, G. A., & Li, C.-H. (2012). Measurement invariance for latent constructs in multiple populations: A critical view and refocus. *Educational and Psychological Measurement*, 72, 954-974.
- Raykov, T., Marcoulides, G. A., & Millsap, R. E. (2013). Examining factorial invariance: A multiple testing procedure. *Educational and Psychological Measurement*, 73, 713-727.
- Raykov, T., Marcoulides, G. A., Lee, C.-L., & Chang, D. C. (2013). Studying differential item functioning via latent variable modeling: A note on a multiple testing procedure. *Educational and Psychological Measurement*, 73, 898-908.
- Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations. *Organizational Research Methods*, 3, 4-70.
- Wasserman, L. W. (2004). *All of statistics*. New York, NY: Springer.
- Woods, C., Cai, L., & Wang, M. (2013). The Langer-improved Wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT. *Educational and Psychological Measurement*, 73, 532-547.