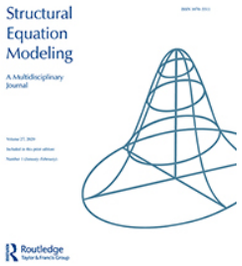
Routledge
Taylor & Francis Group



Structural Equation Modeling: A Multidisciplinary Journal

ISSN: 1070-5511 (Print) 1532-8007 (Online) Journal homepage: https://www.tandfonline.com/loi/hsem20

Simplifying the Assessment of Measurement Invariance over Multiple Background Variables: Using Regularized Moderated Nonlinear Factor Analysis to Detect Differential Item Functioning

Daniel J. Bauer, William C. M. Belzak & Veronica T. Cole

To cite this article: Daniel J. Bauer, William C. M. Belzak & Veronica T. Cole (2020) Simplifying the Assessment of Measurement Invariance over Multiple Background Variables: Using Regularized Moderated Nonlinear Factor Analysis to Detect Differential Item Functioning, Structural Equation Modeling: A Multidisciplinary Journal, 27:1, 43-55, DOI: 10.1080/10705511.2019.1642754

To link to this article: https://doi.org/10.1080/10705511.2019.1642754

Structural Equation Modeling: A Multidisciplinary Journal, 27: 43–55, 2020

ISSN: 1070-5511 print / 1532-8007 online
DOI: https://doi.org/10.1080/10705511.2019.1642754

Simplifying the Assessment of Measurement Invariance over Multiple Background Variables: Using Regularized Moderated Nonlinear Factor Analysis to Detect Differential Item Functioning

Daniel J. Bauer, William C. M. Belzak, and Veronica T. Cole
The University of North Carolina at Chapel Hill

Determining whether measures are equally valid for all individuals is a core component of psychometric analysis. Traditionally, the evaluation of measurement invariance (MI) involves comparing independent groups defined by a single categorical covariate (e.g., men and women) to determine if there are any items that display differential item functioning (DIF). More recently, Moderated Nonlinear Factor Analysis (MNLFA) has been advanced as an approach for evaluating MI/DIF simultaneously over multiple background variables, categorical and continuous. Unfortunately, conventional procedures for detecting DIF do not scale well to the more complex MNLFA. The current manuscript therefore proposes a regularization approach to MNLFA estimation that penalizes the likelihood for DIF parameters (i.e., rewarding sparse DIF). This procedure avoids the pitfalls of sequential inference tests, is automated for end users, and is shown to perform well in both a small-scale simulation and an empirical validation study.

Keywords: Measurement invariance, differential item functioning, moderated nonlinear factor analysis, regularization

The foundation of science is measurement, and the development of both reliable and valid measures has long been a critical enterprise for researchers in psychology and allied fields. One key aspect of validity concerns whether the scores produced by a measure, such as ability scores on an achievement test or depression scores from a symptom inventory, are directly comparable across individuals. If the scores over-estimate the latent trait for some people (e.g., men) and under-estimate it for others (e.g., African-Americans), then observed score differences will not accurately reflect true differences in the quantity being measured (Millsap, 2011). Score comparisons between individuals will be invalid and results obtained from using these scores in subsequent analyses will be distorted. Some effects may be masked, others exaggerated, and still others obtained entirely as artefacts of poor measurement (Curran, Cole, Bauer, Rothenberg, & Hussong, 2018).

Recognizing the importance of this issue, psychometricians have devoted considerable attention to developing theory and methods for assessing whether scores are equivalent in meaning and metric across individuals, a condition referred to as *measurement invariance* (MI). Typically, psychometricians assess measurement invariance by fitting latent variable models to data obtained from multi-item scales, for instance, a depression scale consisting of items measuring sadness, loneliness, crying, etc. To be invariant, the measure must produce the same probability distribution for each item at any specific level of the latent variable, irrespective of the values of any background variables. For instance, the probability of endorsing "crying" should be the same for any two adolescents with the same level of depression, regardless if one is a boy and the other a girl. When a particular item fails to show this property, this indicates *differential item functioning* (DIF). For instance, girls tend to endorse crying more often than boys, even after accounting for overall sex

Correspondence should be addressed to Daniel Bauer, Department of Psychology and Neuroscience, The University of North Carolina at Chapel Hill. E-mail: dbauer@email.unc.edu

Supplemental data for this article can be accessed here.

differences in levels of depression (Steinberg & Thissen, 2006). Failing to accommodate for this DIF would lead one to exaggerate sex differences in depression. When some but not all items on a scale have DIF, this condition is referred to as partial invariance (Byrne, Shavelson, & Muthén, 1989). Partially invariant measures can still yield valid comparisons, but only if DIF has been accurately modeled. A critically important task for psychological measurement is thus to correctly identify which items do and do not display DIF.

A great deal of research has been conducted on methods for evaluating MI/DIF, with proposals involving the application of likelihood ratio tests (Thissen, Steinberg, & Wainer, 1993), Wald tests (Lord, 1980; Woods, Cai, & Wang, 2013), and score tests (modification indices or Lagrange multiplier tests; Oort, 1998; Steenkamp & Baumgartner, 1998). Although exceptions certainly exist, the majority of research on MI/DIF shares two specific limitations. First, traditional approaches typically rely on fitting "multiple groups models" to evaluate MI over levels of a single categorical variable without consideration of other correlates, confounders, or potential moderators. Given the intersectionality of race, sex, and other background variables, this approach is overly simplistic. Second, most procedures for DIF detection rely on significance testing and involve multiple model comparisons. These procedures are tedious to implement, not easily generalized beyond the simple case of comparing independent groups, and resemble forward and backward selection procedures that can lead to suboptimal results (Derksen & Keselman, 1992; MacCallum, Roznowski, & Necowitz, 1992; Whittingham, Stephens, Bradbury, & Freckleton, 2006).

Over the past decade, our research group has developed and advocated for the use of *moderated nonlinear factor analysis* (MNLFA) as an alternative to the multiple groups model for evaluating MI/DIF (Bauer, 2017; Bauer & Hussong, 2009; Curran et al., 2014; Curran, Cole, Bauer, Hussong, & Gottfredson, 2016; Curran et al., 2018; Curran et al., 2018). MNLFA generalizes traditional latent variable models to allow for the simultaneous evaluation of MI/DIF over multiple, correlated background variables, where these may be categorical (e.g., sex), continuous (e.g., age), or interactions (e.g., sex × age). MNLFA thus removes the limitation of considering MI/DIF for only a single grouping variable at a time. The cost of this generalization, however, is increased model complexity and the difficulty of adapting existing DIF detection procedures for use with MNLFA. Indeed, MNLFA compounds all the problems of conventional DIF detection procedures for multiple groups, vastly expanding the ways in which DIF can manifest, increasing the number of potential model comparisons and inference tests, and reducing the likelihood of identifying the correct pattern of DIF.

The present paper proposes to overcome this problem through the implementation of regularization techniques with MNLFA. Specifically, we propose fitting MNLFA using a penalized likelihood function that implements a lasso (least absolute shrinkage and selection operator) penalty on the DIF parameters. This approach has significant advantages relative to conventional DIF detection techniques: it is fully automated, does not depend on inference tests, and relies on a selection technique known to produce more stable results than stepwise selection procedures. In what follows, we explicate this approach to DIF detection, provide initial simulation results on its performance, and further test its validity in an empirical demonstration. First, however, we define MI/DIF more formally, and clarify how it is assessed within the MNLFA.

MEASUREMENT INVARIANCE AND DIFFERENTIAL ITEM FUNCTIONING

MI is said to exist if the distribution of possible item responses for an individual depends only on the person's value for the latent variable and not also on other characteristics of the individual (Mellenbergh, 1989). Following Millsap (2011, p. 46) the mathematical expression of this definition is

$$f(\mathbf{y}_i|\eta_i, \mathbf{x}_i) = f(\mathbf{y}_i|\eta_i) \quad (1)$$

where f designates a probability distribution, $\mathbf{y}_i$ is a $p \times 1$ vector containing the observed item responses for person i, η_i is the unobserved latent factor, and $\mathbf{x}_i$ is a $q \times 1$ vector of observed person-level characteristics (e.g., gender, ethnicity, or age).[1] In words, Equation (1) states that MI exists if the distribution of the observed items depends only on the values of the latent variable. Variables included in $\mathbf{x}$ can relate to η but have no direct influence on the distribution of $\mathbf{y}$ beyond their influence on η. If, for any given item j, this does not hold, i.e.,

$$f(y_{ij}|\eta_i, \mathbf{x}_i) \neq f(y_{ij}|\eta_i) \quad (2)$$

then this represents DIF, indicating that the response distribution for this item differs across individuals as a function of $\mathbf{x}$ beyond what might be expected due to differences on η alone.

Although clearly more general at a theoretical level, the evaluation of MI/DIF in practice has focused almost exclusively on the situation where the covariate vector $\mathbf{x}$ consists of a single grouping variable. This focus began with investigators making informal comparisons of exploratory factor analyses stratified by group. Later, the development of multiple group confirmatory factor analysis by Jöreskog (1971) and Sörbom (1974) permitted formal inferential tests of measurement invariance, again across discrete groups. At essence, this restriction to a grouping variable resulted from a computational approach that relied on

[1] We focus for simplicity on the case of a unidimensional latent variable model; extensions to multidimensional latent variable models are straightforward.

segmenting the population into independent sets (e.g., boys versus girls) and formulating the log-likelihood function as a proportionally weighted sum of the log-likelihoods obtained within each set. This approach, however, does not easily accommodate multiple background variables, requiring that each be considered separately. Neither does it allow for continuous background variables like age or family income, which would have to be discretized. As noted by Bauer (2017), multiple indicator multiple cause (MIMIC) models can address some of these limitations (Woods & Grimm, 2011), but the MNLFA model described below offers a fully general solution.

MODERATED NONLINEAR FACTOR ANALYSIS

Initially proposed by Bauer and Hussong (2009), the MNLFA model offers the ability to evaluate MI/DIF simultaneously across multiple correlated background variables which may be categorical or continuous in nature. Adopting a generalized linear modeling approach, we begin by denoting the expected value of the conditional distribution for the response of person i to item j as $E(y_{ij}) = \mu_{ij}$. This expected value is then expressed as a function of the latent variable via the equation

$$\mu_{ij} = g_j^{-1}(\nu_{ij} + \lambda_{ij}\eta_i) \tag{3}$$

where g denotes a link function chosen in accord with the scale type of the item (e.g., binary or continuous) and which may differ across items, ν denotes an intercept, and λ denotes a slope or factor loading. These intercepts and loadings may vary deterministically across individuals as a function of the covariates as follows,

$$\nu_{ij} = \nu_{j0} + \boldsymbol{\kappa}'_j \mathbf{x}_i \tag{4}$$

$$\lambda_{ij} = \lambda_{j0} + \boldsymbol{\omega}'_j \mathbf{x}_i \tag{5}$$

where the baseline values (where all covariates are zero) are designated with a null subscript and the effects of the covariates on the intercepts and loadings are conveyed by the coefficient vectors $\boldsymbol{\kappa}_j$ and $\boldsymbol{\omega}_j$, respectively.[2] If either of the item parameters differs as a function of the covariates then this signifies the presence of DIF.

[2] If the conditional distribution of the item also includes a dispersion parameter (e.g., for a normally distributed continuous item) then this too may be expressed as a log-linear function of the covariates (Bauer, 2017). For ordinal items, the expected value is expanded to a vector of cumulative probabilities and category specific intercepts or thresholds must be added to the model (Bauer & Hussong, 2009).

The latent factor is assumed to be conditionally normal, with mean $E(\eta_i) = \alpha_i$ and variance $V(\eta_i) = \psi_i$, both of which may also depend on the covariates via the equations:

$$\alpha_i = \alpha_0 + \boldsymbol{\gamma}'\mathbf{x}_i \tag{6}$$

$$\psi_i = \psi_0 \exp(\boldsymbol{\beta}'\mathbf{x}_i) \tag{7}$$

where the baseline values (where all covariates are zero) are again designated with a null subscript and the effects of the covariates are conveyed by the coefficient vectors $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$. Effects of the covariates on the parameters of the latent variable distribution are referred to as *impact* and do not constitute violation of MI (e.g., with increasing age, mathematics ability may truly increase on average and also become more variable across children). Including impact can yield substantively interesting insights and markedly improve score estimation and performance in secondary models, even in the absence of DIF (Curran et al., 2016, 2018).

Whereas in the multiple groups framework MI/DIF is considered in terms of different item parameters across groups, MNLFA re-conceptualizes MI/DIF in terms of moderation of the item parameter valeus, as shown in Equations (4)-(5) (Bauer, 2017). The MNLFA formulation therefore easily admits multiple covariates, including continuous covariates. This way of expressing DIF also facilitates the application of regularization, as shown in the next section. Additionally, estimation of the MNLFA does not rely on computing the likelihood across independent groups but rather on direct computation of the likelihood at the individual record level, conditional on the observed values of the covariates. More specifically, under the assumption of conditional (local) independence of the item responses, the likelihood is given as

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} \int \phi(\eta_i|\mathbf{x}_i) \prod_{j=1}^{P} f(y_{ij}|\eta_i, \mathbf{x}_i) d\eta_i \tag{8}$$

where $\boldsymbol{\theta}$ is a vector of model parameters, consisting of parameters relating to the mean and variance of the latent factor from Equations (6) and (7) as well as parameters governing the conditional distributions of the item responses from Equations (4) and (5). Obtaining maximum likelihood estimates for the MNLFA requires numerical integration over the distribution of the latent factor (e.g., by adaptive quadrature; see Bauer & Hussong, 2009). This is computationally intensive but tractable with a modern personal computer. Despite this difference in the construction of the likelihood function, if the covariate vector is restricted to indicator variables differentiating a set of independent groups, then the MNLFA reduces to a standard multiple groups factor analysis or IRT model, generating

an equivalent likelihood and identical parameter estimates (Bauer, 2017).

REGULARIZED DIF DETECTION WITH MNLFA

Regularization techniques, including lasso (least absolute shrinkage and selection operator) and ridge, were originally proposed as more stable alternatives to stepwise procedures for predictor selection in regression models (Efron, Hastie, Johnstone, & Tibshirani, 2004; Fan & Li, 2001; Friedman, Hastie, & Tibshirani, 2010; Tibshirani, 1996; Tibshirani, Wainwright, & Hastie, 2015). These approaches penalize the loss function (least squares or maximum likelihood) used to estimate the model as a function of the parameter estimates, for instance using an l_1 norm (lasso) or l_2 norm (ridge) or combination of the two (elastic net), thus favoring models in which some regression coefficients attain values of zero or are minimized. Bayesian regularization approaches accomplish a similar goal by implementing priors that shrink some parameters close to zero (e.g., horseshoe or spike and slab priors). Although widely applied with regression models for many years, recognition of the potential advantages of regularization for latent variable models has come only recently (Jacobucci, Grimm, & McArdle, 2016; Pan, Ip, & Dubé, 2017; Sun, Chen, Liu, Ying, & Xin, 2016; Trendafilov & Adachi, 2015; Yuan, Wu, & Bentler, 2011).

Using regularization specifically for DIF detection (Reg-DIF) has previously been considered by Magis, Tuerlinckx, and De Boeck (2015), Tutz and Schauberger (2015), and Huan (2018), demonstrating the feasibility and potential of this approach relative to conventional DIF detection techniques. Here we extend these prior developments by proposing the use of Reg-DIF within the more general MNLFA model. We implement the lasso penalty as this shrinks some parameters precisely to zero, at which point they are removed from the model. To implement the lasso, the log-likelihood function is augmented with a penalty based on the l_1 norm for the DIF parameters (i.e., sum of absolute values). Specifically, the penalized log-likelihood can be expressed as

$$\ell_{lasso}(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) - \tau\left(\|\boldsymbol{\kappa}\|_1 + \|\boldsymbol{\omega}\|_1\right) \quad (9)$$

where $\ell(\boldsymbol{\theta})$ is the unpenalized model log-likelihood, based on Equation (8), τ is a tuning parameter, and $\boldsymbol{\kappa}$ and $\boldsymbol{\omega}$ are vectors of potential DIF effects on intercepts and loadings, respectively, pooled across all items from Equations (4) and (5). Note that only the parameters in $\boldsymbol{\theta}$ which convey DIF are penalized; baseline item parameters and impact parameters are unpenalized.

By penalizing as a function of the sum of absolute values of the DIF parameters, the lasso favors models in which extraneous DIF parameters (those that do not improve the likelihood much) go to zero, particularly as the tuning parameter value increases. If the tuning parameter is set to zero, this nullifies the penalty and conventional maximum likelihood estimates for the MNLFA will be obtained. In contrast, if the tuning parameter is set to a very large value then all DIF parameter estimates will go to zero and no DIF will be detected. How to select the value of the tuning parameter thus becomes an essential question. Typically, the tuning parameter is incremented over a range of values and optimized either by cross validation or considering information criteria (McNeish, 2015). One common procedure is to use K-fold cross-validation to select the tuning parameter value that yields the estimates with greatest stability. However, given the already computationally intensive nature of the MNFLA, and the need to consider a range of possible values for the tuning parameter, conducting the analysis over K folds is burdensome. We therefore follow other recent implementations of regularized latent variable models (e.g., Jacobucci et al., 2016; Tutz & Schauberger, 2015) in using the Bayesian Information Criterion (BIC) to select the tuning parameter value. Thus, we calculate the BIC of the regularized MNLFA for each value of τ and interpret the fitted model at the specific value of τ for which the minimum BIC was obtained.[3] The remaining non-zero DIF parameters indicate which items have DIF. Thus, Reg-DIF via lasso-penalized MNLFA identifies DIF without requiring sequential inference testing (e.g., item-by-item likelihood ratio tests).

An interesting issue of model identification arises given the inclusion of the penalty term. Specifically, the presence of the penalty term permits estimation of models that would otherwise not be identified (Friedman et al., 2010; Tutz & Schauberger, 2015). When using unpenalized maximum likelihood, a minimally identified MNLFA model is one which meets two conditions: (1) The mean and variance of the latent variable are defined to be zero and one, respectively, when $\mathbf{x} = \mathbf{0}$; and (2) At least one item displays no DIF (either intercept or loading) for each covariate in $\mathbf{x}$. In effect, this implies that one must have sufficient prior knowledge to select at least one invariant "anchor item" prior to fitting the regularized MNLFA model. In the presence of a sufficient penalty, however, Condition 2 may no longer be required, as some of the DIF parameters will be shrunken to zero, making the model estimable even without prior selection of anchor items. This possibility is appealing because there is often little substantive motivation for selecting anchor items a priori and choosing incorrectly results in biased tests of DIF for the other items (Kopf, Zeileis, & Strobl, 2015; Wang, 2004; Woods, 2009).

[3] We remove the penalty contribution from $\ell_{lasso}(\boldsymbol{\theta})$ when computing the BIC.

APPLICATION OF REG-DIF WITH SIMULATED DATA

In this section, we present an initial simulation study designed to demonstrate both the feasibility and promise of applying the proposed Reg-DIF procedure with MNLFA.

Data generation

Drawing on our prior simulation work with MNLFA, we selected data from a subset of the conditions described in detail in Curran et al. (2016), (2018)). Population values for the selected study conditions are shown in Table 1. In brief, data generation proceeded in three steps. First, we generated data on three correlated background variables (two binary and one continuous). Inspired by MNLFA applications to facilitate integrative data analysis (i.e., measurement harmonization to enable data pooling across multiple studies), the covariates were simulated to represent *Study* (50% in Study 1, 50% in Study 2), *Gender* (50% male), and *Age* (integer values initially ranging from 9 to 17 then centered at 13). The correlation between gender and study was .30, between age and study was −.51, and between gender and age was −.15. Second, we generated latent variable true scores via Equations (6) and (7), including effects of these covariates on the latent variable. Third, we generated the individual item responses based on Equations (4) and (5). All items were set to be binary, utilizing a Bernoulli response distribution and logistic link function. The form of the MNLFA is then a generalization of the two-parameter logistic model. To observe the primary effects of sample size and number of items (i.e., more information), we examined data simulated under three configurations: N = 500 with 6 items, N = 500 with 12 items (more items), and N = 2000 with 6 items (more people). Within each of these configurations, we simulated data such that either 1/3 or 2/3 of items had DIF. The magnitude of DIF was generated to be either small or large, as determined by weighted area between the curves (Edelen, Stucky, & Chandra, 2015; Hansen et al., 2014).

MNLFA estimation by lasso-penalized maximum likelihood

We implemented Reg-DIF with these data using the NLMIXED procedure within SAS 9.4 (see supplemental materials). As noted by Bauer and Hussong (2009, Appendix), the likelihood in Equation (8) has the same form as the likelihood optimized in NLMIXED, and this can be tailored to incorporate the penalty term in Equation (9). For Reg-DIF analyses, the two binary covariates were effect coded and age was divided by 1.5, yielding population variances of 1.0 for all three covariates. This placed all DIF covariates on the same scale, a necessary feature for implementing regularization (Tibshirani, 1997). We then fit the model with successively larger values of τ, progressively removing DIF parameters from $\boldsymbol{\kappa}$ and $\boldsymbol{\omega}$ that shrunk to zero, and selecting the model with the best BIC. This best-BIC model was then re-fit without the penalty, including only those DIF parameters that had been retained by the regularization procedure (mitigating bias toward zero for the retained DIF parameters due to the penalty). We then considered two ways to identify items with DIF in this final model. The first was to count an item as having DIF if any covariate effects were retained for the item on either the intercept or loading following regularization. The second was to count only items with statistically significant intercept or loading DIF as having DIF (as determined by Wald tests), otherwise the item was counted as not having DIF (either DIF parameters were excluded or they were included but not statistically significant).

TABLE 1
Population Parameter Values for Monte Carlo Simulation Study

	6 Items		*12 Items*		*Intercept (Small DIF | Large DIF)*				*Loading (Small DIF | Large DIF)*			
Item	*DIF 33%*	*DIF 66%*	*DIF 33%*	*DIF 66%*	*Baseline*	*Age*	*Gender*	*Study*	*Baseline*	*Age*	*Gender*	*Study*
1					−.5				1			
2		*	*	*	−.9	.125 | .25	−.5 | −1	.5 | 1	1.3	.05 | .075	−.2 | −.3	.2 | .3
3		*	*	*	−1.3	−.125 | −.25	.5 | 1	.5 | 1	1.6	−.05 | −.075	.2 | .3	.2 | .3
4	*	*	*	*	−1.7	.125 | .25			1.9	.05 | .075		
5	*	*	*	*	−2.1		−.5 | −1	.5 | 1	2.2		−.2 | −.3	.2 | .3
6					−2.5				2.5			
7					−.5				1			
8				*	−.9	.125 | .25	−.5 | −1	.5 | 1	1.3	.05 | .075	−.2 | −.3	.2 | .3
9				*	−1.3	−.125 | −.25	.5 | 1	.5 | 1	1.6	−.05 | −.075	.2 | .3	.2 | .3
10				*	−1.7	.125 | .25			1.9	.05 | .075		
11				*	−2.1		−.5| −1	.5 | 1	2.2		−.2 | −.3	.2 | .3
12					−2.5				2.5			

Note. Asterisks indicate items with DIF in the indicated condition.

Using a high-performance research computing cluster, we ran 100 replications for each of the twelve study conditions, with approximately 100 values of τ per replication, totaling approximately 120,000 estimated models. Values of τ were incremented in a range from 2×10^{-5} to 10^{-3}, with small increment values specified at the beginning of the regularization procedure to ensure DIF parameters were excluded one by one from the active set, and large increment values specified at the end to ensure all parameters had been excluded from the active set. As noted above, the lasso penalty can enable estimation of otherwise under-identified models (functioning somewhat similarly to a prior). Thus, we did not specify an a priori anchor item to identify the model (i.e., all items have potential DIF). As the penalty term grows in magnitude, DIF parameters are removed from the model, at which point the model becomes analytically identified through the presence of anchor items. Convergence/completion rates for Reg-DIF were 97–100% in all conditions except N = 500 with 12 items and DIF on 2/3 of the items, where convergence rates were 86% (large-magnitude DIF) and 88% (small-magnitude DIF).

Benchmark comparison

To benchmark the performance of Reg-DIF, we also computed results using an adaptation of the item response theory likelihood ratio DIF detection procedure (IRT-LR-DIF; Thissen et al., 1993). For this approach, MNLFA models were fit (without any penalty) using Mplus version 8 (Muthén & Muthén, 2017; see Bauer, 2017, for example scripts). First, a baseline model was fit including no DIF on any items. Then, a sequence of models was fit allowing DIF for each covariate on both the intercept and slope of one item at a time, assuming no DIF for any other items (all other items as treated as anchors). When a likelihood ratio test (LRT) indicated that including DIF effects for a given item produced a significant improvement in model fit, then this item was identified as having DIF, otherwise it was counted as not having DIF. Following common practice, the Benjamini-Hochberg procedure was employed in an effort to control the false discovery rate given multiple testing (Benjamini & Hochberg, 1995; Thissen, Steinberg, & Kuang, 2002).

Results

Given that the core goal of Reg-DIF is to identify which items do and do not have DIF, the primary outcomes of interest are item-level false positive (FP) and true positive (TP) rates, corresponding to Type I errors and power, respectively.

Figure 1 displays the false positive rate for Reg-DIF (requiring significance or not for retained DIF parameters) relative to IRT-LR-DIF. As can be seen, IRT-LR-DIF has a much higher false positive rate across all conditions and especially when DIF is pervasive (66%) and of large magnitude. In contrast, Reg-DIF maintains a much lower false positive rate, particularly if DIF is counted only when covariate effects on item parameters are both retained and statistically significant. Each DIF detection approach showed minor improvements in false positive rates with more items. IRT-LR-DIF showed high sensitivity to sample size, with higher Type I error rates at N = 2000. Reg-DIF was insensitive to sample size when the DIF present was large in magnitude, but showed some elevation in false positive rates at large N when DIF was small. Overall, the results indicate that, under the conditions studied here, IRT-LR-DIF has excessively high Type I errors (consistent with Finch, 2005; Millsap, 2011, p. 199–200; Stark, Chernyshenko, & Drasgow, 2006), whereas Reg-DIF is much less likely to identify DIF where it is not.

The true positive rates are displayed in Figure 2. Here, IRT-LR-DIF showed a consistent advantage over Reg-DIF in identifying DIF when it truly is present. In light of the high false positive rates, however, this advantage is offset by the tendency of IRT-LR-DIF to identify DIF indiscriminately. With Reg-DIF, true positive rates increase with effect size (large DIF) and sample size but less so with number of items. When 33% of items had large-magnitude DIF, Reg-DIF generated true positive rates of around .80 (conventionally accepted power levels) at N = 500 with either 6 or 12 items. When 66% of items had DIF or DIF was small in magnitude, however, the larger sample size of N = 2000 was required to reach this level. Requiring that covariate effects on the item parameters both be present and significant reduced true positive rates (but also false positive rates, as discussed above).

Supplementing these results, we also conducted two secondary analyses. First, we examined the false and true positive rates of Reg-DIF when using Akaike's Information Criterion (AIC) for model selection instead of BIC, as shown in Figures 3 and 4. Use of the more liberal AIC results in retention of more DIF parameters by Reg-DIF and thus an elevation of both true and false positives relative to the BIC, particularly when the retained parameters are not required to be significant. Indeed, when not requiring significance, Reg-DIF using AIC performs similarly to IRT-LR-DIF, with high true positive rates mitigated by concomitantly high false positive rates. When requiring the estimates to be significant, the false positive rate elevation of Reg-DIF by AIC relative to BIC was far more muted. The true positive rate increased with AIC relative to BIC most notably only for the N = 500 conditions (when the BIC penalty is stongest) and when small DIF was present for 33% of items. Thus, if DIF is expected to limited to a smaller portion of items and small in magnitude and the sample size is not large, AIC may be preferable to BIC but only if DIF effects are required to both be retained and significant, otherwise the false positive rate is excessive.

Second, we examined false and true positive rates for intercept DIF versus loading DIF parameters (again using BIC for tuning Reg-DIF). Because these analyses were secondary, we provide only a brief summary here

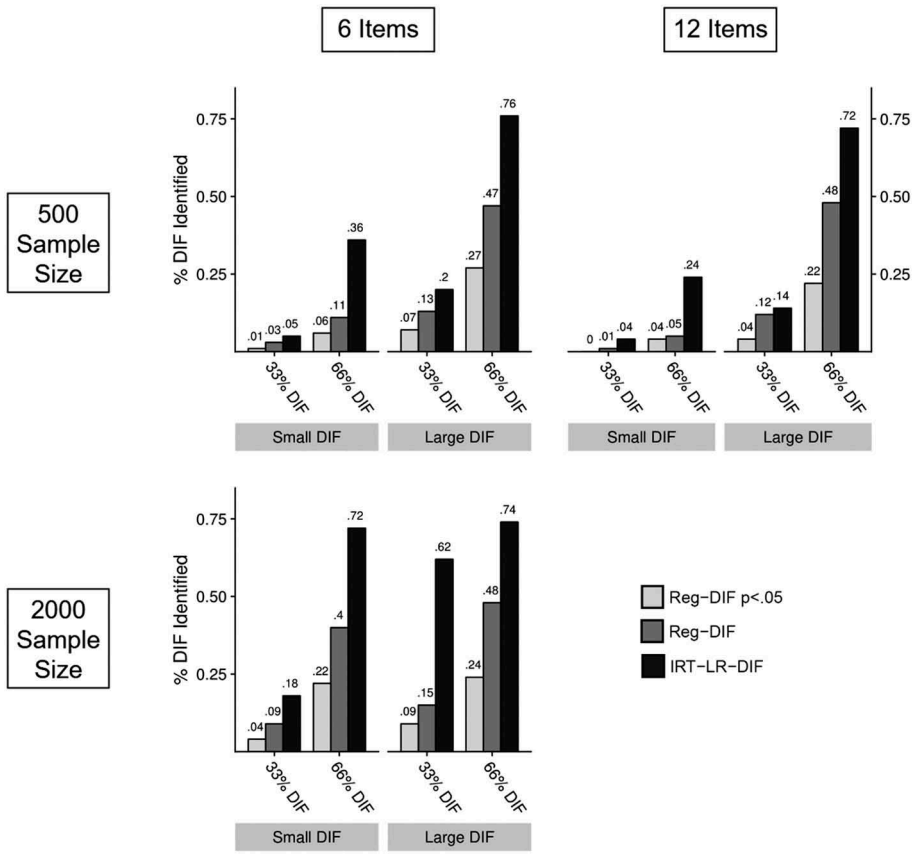


FIGURE 1 Percentage of items without Differential Item Functioning (DIF) for which DIF was (a) retained and significant following regularization (light gray); (b) retained following regularization, regardless of significance (medium gray); (c) significant by the item response theory likelihood ratio test (black). The tuning parameter for regularized DIF evaluation was determined by minimizing Bayes' Information Criterion.

(see supplemental materials). Reg-DIF false and true positive rates for intercept DIF mirrored the item-level rates presented above but were slightly lower. Reg-DIF had more difficulty identifying loading DIF. When requiring both retention and significance of loading DIF, false positive rates were generally at or below 5%, but true positive rates were also low, reaching a maximum of only 26% for N = 2000 with large DIF on 66% of items. IRT-LR-DIF followed by Wald tests on the loading DIF estimates also had low power, although higher than Reg-DIF. The highest true positive rate for loading DIF was achieved by Reg-DIF without requiring significance, but this came at the cost of unacceptably large false positive rates. The general pattern was that all procedures were most sensitive to intercept DIF.

Summary

Taken together, these results indicate that Reg-DIF maintains far better false positive control than IRT-LR-DIF (as adapted for MNLFA). Requiring both retention and significance of DIF parameters in Reg-DIF resulted in the best control of false positive rates but also entailed a loss in power. This approach performed best at identifying items with and without DIF when DIF is large in magnitude and not too pervasive. When DIF is small in magnitude, a much larger sample size is required to adequately detect DIF. Additionally, when the majority of items have DIF, the false positive rate of Reg-DIF may be elevated. Ancillary analyses showed that using BIC to optimize Reg-DIF is preferable to AIC under most conditions except when attempting to identify small DIF effects in smaller samples, and that Reg-DIF is better able to identify DIF with respect to item intercepts than loadings. Overall, using the BIC to optimize the tuning parameter and requiring significance of the estimated DIF parameters appears to offer the best balance of Type I and II errors, but this approach is most likely to be useful in large samples where power is less of a concern. We now demonstrate and further validate the use of Reg-DIF via an empirical application.

EMPIRICAL DEMONSTRATION AND VALIDATION

In this section, we further evaluate Reg-DIF by applying it and IRT-LR-DIF (for comparison) to data from the

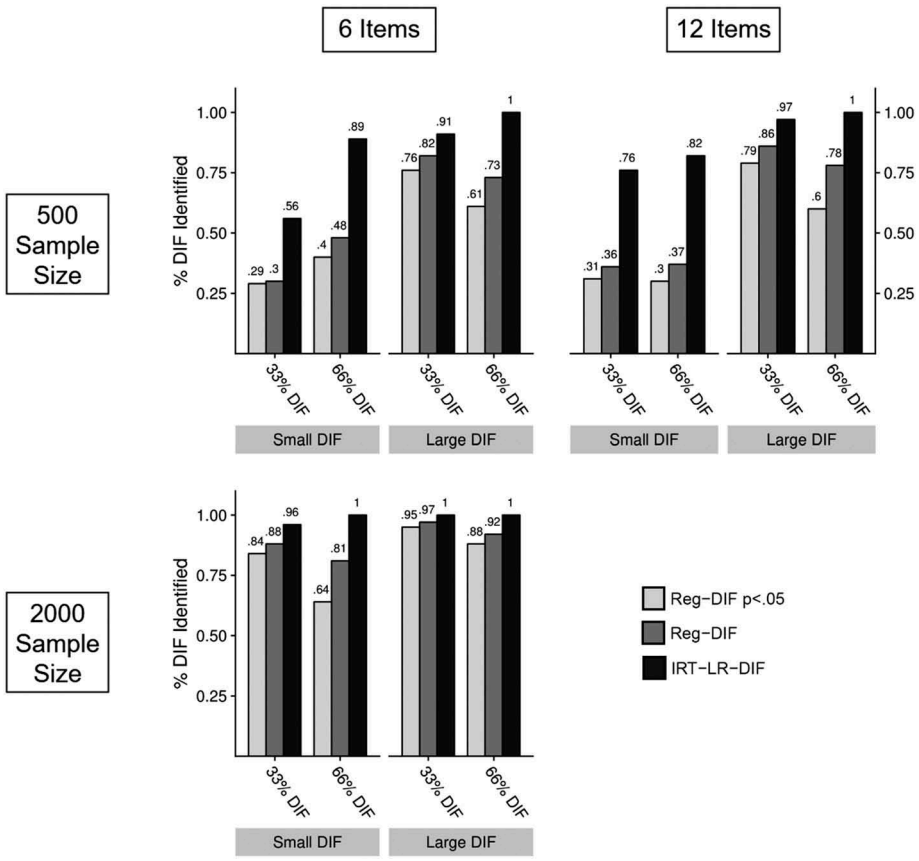


FIGURE 2 Percentage of items with Differential Item Functioning (DIF) for which DIF was (a) retained and significant following regularization (light gray); (b) retained following regularization, regardless of significance (medium gray); (c) significant by the item response theory likelihood ratio test (black). The tuning parameter for regularized DIF evaluation was determined by minimizing Bayes' Information Criterion.

REAL-U project, a study specifically designed to generate human-subjects data with which to empirically validate psychometric methodology. Participants in REAL-U were randomly assigned to different "studies" using similar but somewhat altered measures related to alcohol and substance use. The measure of focus is a shortened form of the Rutgers Alcohol Problem Index (RAPI; White & Labouvie, 1989) that retains 18 of the original 23 items (Neal, Corbin, & Fromme, 2006). In the conditions examined here, the wording of specific items was modified to reflect differences commonly observed across independently conducted studies. For instance, the item "Relatives avoided you" became "Family members rejected you because of your drinking." Given the random assignment of participants to conditions, we would expect only these modified items (and perhaps only a subset) to evince DIF as a function of study.[4] If a DIF detection procedure identifies study differences in only these items, then this provides empirical evidence of the effectiveness of the technique.

In fitting the models, we also simultaneously allowed for potential gender DIF, both to illustrate the generality of the MNLFA and because prior research has found gender DIF for some items on the 18-item RAPI. In particular, using a large sample of college student drinkers, Earleywine, LaBrie, and Pedersen (2008) found gender DIF with respect to the intercepts of Items 2, 13, 14, and 15. They did not evaluate possible loading DIF.

Sample and items

The sample included 854 subjects (54% female) between ages 18 and 23 who were enrolled at a large Southeastern university and indicated past-year alcohol use. On the RAPI, participants were asked to "Indicate how many times each of the following things happened to you at some point in your life." Each item is a cognitive or behavioral consequence that occurred while

[4] This assumes there are no context effects, wherein changing the wording of an item influences the interpretation of other items (even if identically worded). Under this assumption, only modified items would be expected to show DIF by study. Changing the wording of an item stem increases the likelihood of DIF but does not guarantee that DIF will occur, thus only a subset of the modified items may ultimately display DIF.

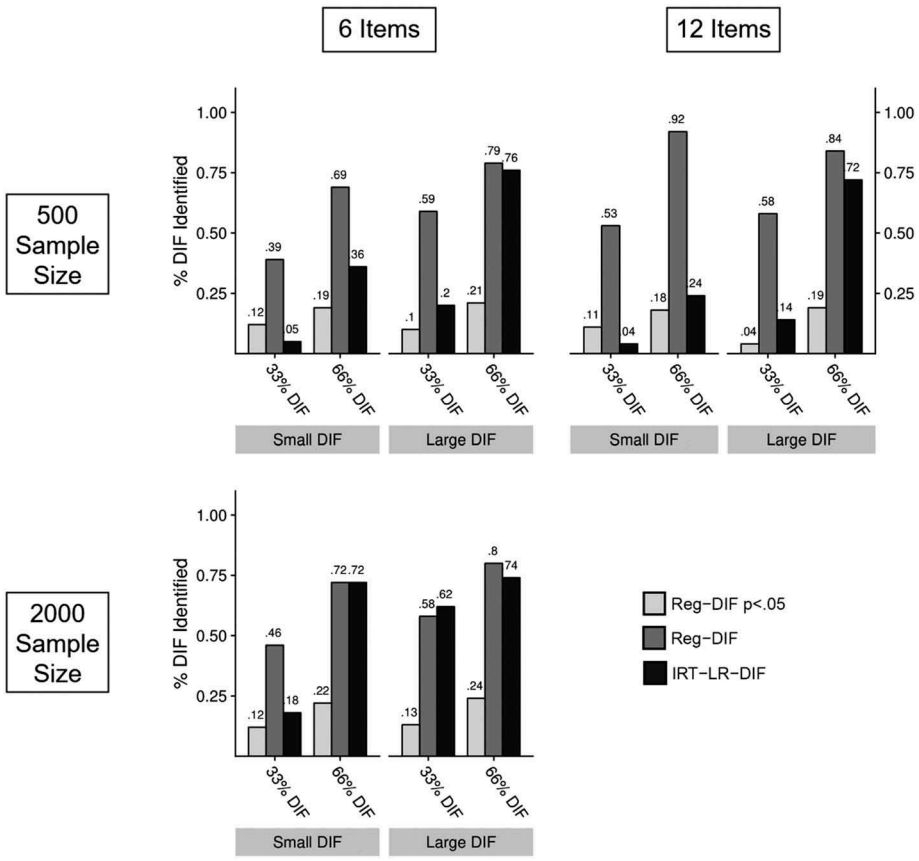


FIGURE 3 Percentage of items without Differential Item Functioning (DIF) for which DIF was (a) retained and significant following regularization (light gray); (b) retained following regularization, regardless of significance (medium gray); (c) significant by the item response theory likelihood ratio test (black). The tuning parameter for regularized DIF evaluation was determined by minimizing Akaike's Information Criterion.

drinking or because of drinking alcohol. Item responses were ordinal, ranging from 0 to 3 with anchors as follows: None (0), 1–2 times (1), 3–5 times (2), More than 5 times (3), or Refuse to answer (missing). On average, 77% of the responses fell into the None (0) category. We thus collapsed item responses into two options: Not present (0) or Present (1). Approximately half of the participants (414) were given the RAPI items as usual, and the other half (440) were given a perturbed version of the RAPI, with altered item stems for 9 of the 18 items. Table 2 displays the original and altered items.

Results

The results obtained from applying Reg-DIF (best-BIC method) are displayed in Table 3. Reg-DIF identified study DIF for four of the nine perturbed items. In particular, Items 2, 9, 12, and 13 were perturbed across studies and also identified as having study DIF. Items 9, 12, and 13 evinced statistically significant differences in the item intercepts between "study" groups, while Item 2 showed non-significant DIF in both intercepts and slopes. Only one unperturbed item, Item 1, was identified as having study DIF. Additionally, the regularization procedure included gender DIF parameters for Items 14, 15, and 16. These differences were statistically significant for Item 14 (intercept) and Item 16 (slope), but not Item 15. Earleywine et al. (2008) also identified gender DIF in Items 14 and 15 but did not evaluate possible loading DIF as found for Item 16. Reg-DIF did not identify gender DIF in either Item 2 or 13.

For comparison, we also ran the adapted IRT-LR-DIF procedure, which identified DIF in ten items, including six perturbed items (Items 2, 3, 9, 12, 13, and 16) and four unperturbed items (Items 1, 4, 14, and 15). Inspection of the univariate Wald tests for the unperturbed items indicated that significant study DIF was detected for Item 1 (intercept DIF), 4 (intercept and loading), and 14 (intercept). Of the four items identified by Earleywine et al. (2008) to have gender DIF, significant gender DIF was found for Items 13 and 14 (intercept) but not Items 2 or 15. For Item 2, only Study DIF was significant. For Item 15, none of the individual DIF parameter estimates was significant, despite the significant joint LRT, thus the source of DIF was ambiguous.

Discussion

The results obtained from this real-world empirical validation study are consistent with the simulation study. In comparison

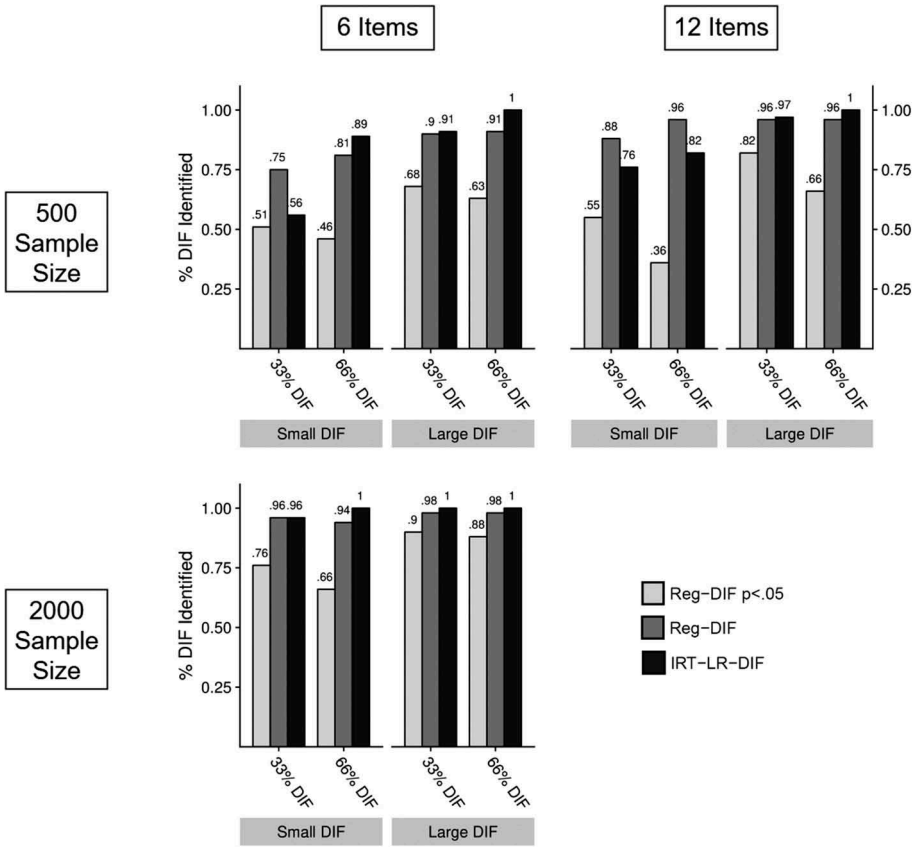


FIGURE 4 Percentage of items with Differential Item Functioning (DIF) for which DIF was (a) retained and significant following regularization (light gray); (b) retained following regularization, regardless of significance (medium gray); (c) significant by the item response theory likelihood ratio test (black). The tuning parameter for regularized DIF evaluation was determined by minimizing Akaike's Information Criterion.

TABLE 2
Original and Perturbed Items for the Rutgers Alcohol Problems Index (RAPI)

Item	*Original Item Prompt*	*Pertubed Item Prompt*
1	Got into fights with other people (friends, relatives, strangers)	
2	Went to work or school high or drunk	Gone to class or a job when drunk
3	Caused shame or embarrassment to someone	Made others ashamed by your drinking behavior or something you did when drinking
4	Neglected your responsibilities	
5	Relatives avoided you	Family members rejected you because of your drinking
6	Felt that you needed more alcohol than you used to in order to get the same effect	
7	Tried to control your drinking (tried to drink only at certain times of the day or in certain places, that is, tried to change your pattern of drinking)	
8	Had withdrawal symptoms, that is, felt sick because you stopped or cut down on drinking	
9	Noticed a change in your personality	Acted in a very different way or did things you normally would not do because of your drinking
10	Felt that you had a problem with alcohol	
11	Wanted to stop drinking but couldn't	Tried unsuccessfully to stop drinking
12	Suddenly found yourself in a place that you could not remember getting to	Awakened the morning after some drinking the night before and could not remember a part of the evening.
13	Passed out or fainted suddenly	Passed out after drinking
14	Had a fight, argument, or bad feeling with a friend	
15	Kept drinking when you promised yourself not to	
16	Felt you were going crazy	Your drinking made you feel out of control even when you were sober
17	Felt physically or psychologically dependent on alcohol	
18	Was told by a friend, neighbor or relative to stop or cut down drinking	Near relative or close friend worried or complained about your drinking

TABLE 3
Differential Item Functioning (DIF) Identified by the Regularized-DIF Procedure

	Covariate for which DIF was identified	
Item	*Intercept DIF*	*Loading DIF*
1	**Study**	
2	Study	Study
9	**Study**	
12	**Study**	
13	**Study**	
14	**Gender**	
15	Gender	Gender
16		**Gender**

Note. Bold entries are statistically significant

to the adapted IRT-LR-DIF procedure, Reg-DIF identified study DIF in two fewer perturbed items but also identified study DIF in only one item that was administered identically across samples, whereas IRT-LR-DIF identified DIF in four identical items, with three showing significant study DIF. No Study DIF was expected for the identical items, and thus the higher number of these items identified to have Study DIF by IRT-LR-DIF is most likely a reflection of the elevated Type I error rate of this procedure. Results for gender DIF were roughly comparable between the two procedures. Of the four items previously identified by Earleywine et al. (2008) to have gender DIF, both Reg-DIF and IRT-LR-DIF identified two as having significant gender DIF. Reg-DIF also retained non-significant gender DIF for another of these four items. IRT-LR-DIF identified DIF for the both of the remaining two items but follow-up Wald tests did not localize this to gender in either case. Finally, Reg-DIF identified significant slope DIF as a function of gender on one additional item, a type of DIF not tested by Earleywine et al. (2008).

Overall, the two procedures were best differentiated with respect to identifying study DIF where the experimental design also permits stronger conclusions regarding accuracy. Results confirm that IRT-LR-DIF tends to find DIF both where it should and where it shouldn't, whereas Reg-DIF provides a better balance of Type I and II errors.

CONCLUSIONS AND FUTURE DIRECTIONS

The key advantage of MNLFA, relative to many other psychometric models, is that it allows one to examine DIF simultaneously across multiple, potentially correlated, covariates. Yet the corresponding difficulty of using MNLFA is the sheer number of potential DIF parameters that could be included in the model and how best to determine which of these to include on the basis of the empirical information at hand. Consider the scope of the problem in our simulation study. With six items, two item parameters per item, and three covariates, there are a total of 36 possible DIF parameters to include in the model. With twelve items, this rises to 72 possible DIF parameters. Likewise, in the REAL-U empirical analysis there were 18 items, two item parameters per item, and two covariates, also yielding 72 possible DIF parameters. Previous efforts at DIF detection using MNLFA have relied on extensions of existing approaches for DIF detection in standard multi-group models (e.g., Bauer & Hussong, 2009; Bauer, 2017; Curran et al., 2014), such as the IRT-LR-DIF procedure considered here. Yet such extensions are unappealing for several reasons. First, they involve specifying and contrasting the fit of many models, a time consuming and error-prone process that requires high expertise (and patience) by the user. Second, the sequential likelihood ratio tests often used to identify DIF are unlikely to adhere to their test distributions (Maydeu-Olivares & Cai, 2006) and risk excessive Type I errors (as shown here and by Finch, 2005; Millsap, 2011, p. 199–200; Stark et al., 2006). Third, how best to extend these techniques in the presence of multiple covariates is unclear – should tests be conducted itemwise across covariates (as implemented here), or by covariate across items, or for each individual DIF parameter?

The Reg-DIF procedure advanced here offers a more appealing strategy for taming the complexity of DIF detection in MNLFA models. Without recourse to inference tests, Reg-DIF uses a penalized estimation approach to arrive at a model in which DIF is included only for those covariates and only on those items where its inclusion meaningfully increases ability of the model to reproduce the observed item responses. Reg-DIF is also a naturally automated procedure (see supplemental materials demonstrating how to run Reg-DIF in SAS), easing the burden on end users who might otherwise need to fit and compare many models manually (making subjective decisions along the way regarding both the ordering and interpretation of the model comparisons). Finally, Reg-DIF performed well in the simulations and empirical validation reported here. Our results indicate that the Reg-DIF procedure can effectively identify DIF where DIF is present and avoid identifying DIF where it is not. It performs best at identifying true DIF when DIF is large in magnitude and the sample size is large and is more likely to identify DIF erroneously when DIF is particularly pervasive. These qualifications are unsurprising: It is hard to imagine any DIF detection procedure that would not be similarly affected by sample size and extent of DIF in a scale.

Although the results presented here for Reg-DIF are encouraging, they are also subject to a number of limitations. The scope of our initial simulation study was small and it will be important in future research to conduct larger simulations to more clearly delimit the data conditions under which Reg-DIF performs well. To enable more systematic simulations, however, it will also be necessary to implement the Reg-DIF procedure with greater computational efficiency, as the NLMIXED procedure we used is particularly time intensive. In our experience, Mplus is much more computationally efficient for fitting unpenalized MNLFA models, so it follows that it should also be

possible to increase the computational efficiency of fitting penalized MNLFA models. Computational developments must also attend to the fact that standard optimization techniques can have difficulty finding an optimal solution for lasso-penalized likelihood functions (Yuan, Chang, Hsieh, & Lin, 2010), a problem that we sought (somewhat clumsily) to circumvent in our implementation of NLMIXED by removing parameters from the model as they attained values of approximately zero. Increasing the speed and stability of the Reg-DIF procedure would also add to its appeal for applied researchers. Last, there are many potential extensions to the lasso regularization approach used here that are worthy of future consideration. For instance, in the current implementation, we applied the same tuning parameter value to all DIF parameters, irrespective of the item, covariate, or item parameter (intercept or loading). An adaptive lasso implementation would permit differential penalization of parameters, perhaps increasing sensitivity to loading DIF. Additionally, we implemented Reg-DIF within a frequentist framework through a penalized likelihood function, and it would be worthwhile to explore the potential advantages of using a Bayesian regularization approach as an alternative (e.g., Brandt, Cambria, & Kelava, 2018). While acknowledging these limitations and the need for future research, we nevertheless believe that the current findings provide strong preliminary evidence in favor of Reg-DIF as a procedure for evaluating MI and detecting DIF in MNLFA models.

ACKNOWLEDGMENTS

he content is solely the responsibility of the authors and does not represent the official views of the National Institute on Drug Abuse or the National Institutes of Health.

FUNDING

This work was supported by the National Institutes of Health [R01 DA034636]. (PI: Daniel Bauer)

ORCID

William C. M. Belzak http://orcid.org/0000-0001-6594-1651

REFERENCES

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, *22*, 507–526. doi:10.1037/met0000077

Bauer, D. J., & Hussong, A. M. (2009). Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. *Psychological Methods*, *14*, 101–125. doi:10.1037/a0015583

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.

Brandt, H., Cambria, J., & Kelava, A. (2018). An adaptive Bayesian lasso approach with spike-and-slab priors to identify multiple linear and nonlinear effects in structural equation models. *Structural Equation Modeling*, *25*, 946–960, doi:10.1080/10705511.2018.1474114

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*, 456–466. doi:10.1037/0033-2909.105.3.456

Curran, P. J., Cole, V., Bauer, D. J., Hussong, A. M., & Gottfredson, N. (2016). Improving factor score estimation through the use of observed background characteristics. *Structural Equation Modeling*, *23*, 827–844. doi:10.1080/10705511.2016.1220839

Curran, P. J., Cole, V. T., Bauer, D. J., Rothenberg, A. W., & Hussong, A. M. (2018). Recovering predictor-criterion relations using covariate-informed factor score estimation. *Structural Equation Modeling, 25,* 860–875. doi:10.1080/10705511.2018.1473773

Curran, P. J., McGinley, J. S., Bauer, D. J., Hussong, A. M., Burns, A., Chassin, L., … Zucker, R. (2014). A moderated nonlinear factor model for the development of commensurate measures in integrative data analysis. *Multivariate Behavioral Research*, *49*, 214–231. doi:10.1080/00273171.2014.889594

Derksen, S., & Keselman, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, *45*, 265–282. doi:10.1111/bmsp.1992.45.issue-2

Earleywine, M., LaBrie, J. W., & Pedersen, E. R. (2008). A brief Rutgers Alcohol Problem Index with less potential for bias. *Addictive Behaviors*, *33*, 1249–1253. doi:10.1016/j.addbeh.2008.05.006

Edelen, M. O., Stucky, B. D., & Chandra, A. (2015). Quantifying "problematic" DIF within an IRT framework: Application to a cancer stigma index. *Quality of Life Research*, *24*, 95–103. doi:10.1007/s11136-013-0540-4

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, *32*, 407–499. doi:10.1214/009053604000000067

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348–1360. doi:10.1198/016214501753382273

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

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*, 1–22.

Hansen, M., Cai, L., Stucky, B. D., Tucker, J. S., Shadel, W. G., & Edelen, M. O. (2014). Methodology for developing and evaluating the PROMIS® smoking item banks. *Nicotine & Tobacco Research*, *16*, S175–S189. doi:10.1093/ntr/ntt123

Huan, P.-H. (2018). A penalized likelihood method for multi-group structural equation modeling. *British Journal of Mathematical and Statistical Psychology*, *71*, 499–522. doi:10.1111/bmsp.12130

Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized structural equation modeling. *Structural Equation Modeling*, *23*, 555–566. doi:10.1080/10705511.2016.1154793

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*, 409–426. doi:10.1007/BF02291366

Kopf, J., Zeileis, A., & Strobl, C. (2015). Anchor selection strategies for DIF analysis: Review, assessment, and new approaches. *Educational and Psychological Measurement*, *75*, 22–56. doi:10.1177/0013164414529792

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

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*, 490–504.

Magis, D., Tuerlinckx, F., & De Boeck, P. (2015). Detection of differential item functioning using the lasso approach. *Journal of Educational and Behavioral Statistics*, *40*, 111–135. doi:10.3102/1076998614559747

Maydeu-Olivares, A., & Cai, L. (2006). A cautionary note on using G^2 (dif) to assess relative model fit in categorical data analysis. *Multivariate Behavioral Research*, *41*, 55–64. doi:10.1207/s15327906mbr4101_4

McNeish, D. M. (2015). Using lasso for predictor selection and to assuage overfitting: A method long overlooked in behavioral sciences. *Multivariate Behavioral Research*, *50*, 471–484. doi:10.1080/00273171.2015.1036965

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*, 127–143. doi:10.1016/0883-0355(89)90002-5

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Muthén, L. K., & Muthén, B. O. (2017). *Mplus User's Guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.

Neal, D. J., Corbin, W. R., & Fromme, K. (2006). Measurement of alcohol-related consequences among high school and college students: Application of item response models to the Rutgers alcohol problem index. *Psychological Assessment*, *18*, 402–414. doi:10.1037/1040-3590.18.4.402

Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling*, *5*, 107–124. doi:10.1080/10705519809540095

Pan, J., Ip, E. H., & Dubé, L. (2017). An alternative to post hoc model modification in confirmatory factor analysis: The Bayesian lasso. *Psychological Methods*, *22*, 687–704. doi:10.1037/met0000112

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology*, *27*, 229–239. doi:10.1111/bmsp.1974.27.issue-2

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, *91*, 1292–1306. doi:10.1037/0021-9010.91.6.1292

Steenkamp, J. B. E., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*, 78–90. doi:10.1086/209528

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. *Psychological Methods*, *11*, 402–415. doi:10.1037/1082-989X.11.4.402

Sun, J., Chen, Y., Liu, J., Ying, Z., & Xin, T. (2016). Latent variable selection for multidimensional item response theory models via a L1 regularization. *Psychometrika*, *81*, 921–939. doi:10.1007/s11336-016-9529-6

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

Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of Educational and Behavioral Statistics*, *27*, 77–83. doi:10.3102/10769986027001077

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, *58*, 267–288.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine*, *16*, 385–395.

Tibshirani, R., Wainwright, M., & Hastie, T. (2015). *Statistical learning with sparsity: The lasso and generalizations*. Boca Raton, FL: Chapman and Hall/CRC.

Trendafilov, N. T., & Adachi, K. (2015). Sparse versus simple structure loadings. *Psychometrika*, *80*, 776–790. doi:10.1007/s11336-014-9416-y

Tutz, G., & Schauberger, G. (2015). A penalty approach to differential item functioning in Rasch models. *Psychometrika*, *80*, 21–43. doi:10.1007/s11336-013-9377-6

Wang, W.-C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *Journal of Experimental Education*, *72*, 221–261. doi:10.3200/JEXE.72.3.221-261

White, H. R., & Labouvie, E. W. (1989). Towards the assessment of adolescent problem drinking. *Journal of Studies on Alcohol*, *50*, 30–37.

Whittingham, M. J., Stephens, P. A., Bradbury, R. B., & Freckleton, R. P. (2006). Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology*, *75*, 1182–1189. doi:10.1111/j.1365-2656.2006.01141.x

Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, *33*, 42–57. doi:10.1177/0146621607314044

Woods, C. M., 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. doi:10.1177/0013164412464875

Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement*, *35*, 339–361. doi:10.1177/0146621611405984

Yuan, G.-X., Chang, K.-W., Hsieh, C.-J., & Lin, C.-J. (2010). A comparison of optimization methods and software for large-scale L1-regularized linear classification. *Journal of Machine Learning Research*, *11*, 3183–3234.

Yuan, K. H., Wu, R., & Bentler, P. M. (2011). Ridge structural equation modelling with correlation matrices for ordinal and continuous data. *British Journal of Mathematical and Statistical Psychology*, *64*, 107–133. doi:10.1348/000711010X497442