Comparison of Measurement Invariance Testing using Penalized Likelihood and Maximum Likelihood Estimators: A Monte Carlo Simulation Study

W. Holmes Finch
Ball State University

Invariance testing remains a widely used and important issue for social scientists. At its heart, assessment of factor invariance involves an examination of the suitability of a scale's use across an entire population. Traditionally, invariance testing has been carried out using a Chi-square difference test in conjunction with multiple group confirmatory factor analysis. However, research has demonstrated that this approach can result in inflated Type I error rates, or findings of a lack of invariance when in fact invariance is present. As a result, statisticians and methodologists have been investigating alternative approaches to testing invariance, which control the Type I error rate without sacrificing much in terms of power. The current study investigated one such alternative, based on a penalized likelihood estimator. This estimator has been previously investigated in the context of fitting structural equation models, and found to perform well in terms of parameter estimation accuracy. Results of the current Monte Carlo simulation study found that the PLE approach is in fact promising in the context of invariance assessment. It was able to control the Type I error rate better than did the Chi-square test, and it exhibited power rates that were as good as or better than those of the Chi-square. Implications of these findings are discussed.

The invariance of latent variable models is an important issue in a wide variety of fields within the social sciences. Invariance refers to the case where latent variable model parameters, such as factor loadings, factor intercepts, or error variances, are equivalent across subgroups within the population. It is key for users of educational and psychological scales, as its presence allows for the use of such instruments with the entire population of interest. On the other hand, when invariance cannot be demonstrated, users of the scale cannot be certain that scores produced by it have the same meaning across subgroups, such as different ethnic groups, genders, or individuals with different socioeconomic status (Dorans, & Cook, 2016; Millsap, 2011; Wu, Li, & Zumbo, 2007). Thus, researchers who do plan to use scales with broad populations of individuals need to demonstrate scale invariance.

The investigation of latent trait model parameter invariance typically involves the use of multiple groups confirmatory factor analysis (MGCFA). In this paradigm, the fit of models with, and without group equality constraints on the model parameters are compared, and if the fit of the models differs, we conclude that invariance does not hold (Millsap, 2011). Perhaps the most common statistical approach used in such invariance assessment involves the calculation of the Chi-square difference statistic, which is discussed in more detail below. However, research has demonstrated that in some situations, this approach has an inflated Type I error rate, resulting in a rejection of the null hypothesis of invariance when in fact invariance holds within the population (Yuan & Bentler, 2004). The purpose of the current study is to examine the performance of an invariance assessment approach based upon the use of a penalized likelihood estimator (PLE) for latent variable models (Huang, 2018), and which might prove to be a worthy alternative to the chi-square difference based approach. The paper is organized as follows. First, a brief review of the MGCFA approach to testing factor invariance (FI) is presented. Next, PLE is discussed, followed by a description of how it can be used to assess FI. The goals of the study, including research questions and hypotheses are then presented, as is the methodology used to address them. Finally, the results of the simulation study and a discussion of those results are presented.

Factor Invariance Assessment with MGCFA

The general factor model takes the following form:

$$x = \tau + \Lambda\xi + \delta \quad (1)$$

where x = Vector of observed indicator variables; e. g. items on a scale or subscale scores
ξ = Vector of latent traits being measures by x
Λ = Matrix of factor loadings linking x and ξ
τ = Vector of intercepts associated with x
δ = Vector of unique errors associated with x

The model in (1) implies the following covariance matrix for the observed indicators:

$$\Sigma = \Lambda\Psi\Lambda' + \Theta \tag{2}$$

where $\Sigma =$ Covariance matrix of the observed indicators, x
$\Psi =$ Covariance matrix of the latent factors
$\Theta =$ Covariance matrix of unique error terms, assumed to be diagonal

As noted above, FI occurs when the parameters in equations (1) and (2) are equivalent across subgroups within the population. There exist different levels of FI, each of which makes different assumptions about the nature of such equivalencies. The weakest type of invariance is referred to as configural invariance (CI), for which it is assumed that only the basic factor structure (i.e., the number of latent variables and the correspondence of observed indicators to these variables) is the same across groups. If CI is present, researchers typically next assess whether the factor loadings (Λ) are equal across groups, in what is known as measurement invariance (MI). In turn, when MI is present, the equality of the factor model intercepts (τ) across groups is next determined, for what is known as structural invariance (SI). Finally, if SI holds, the researcher can assess whether there is group invariance with respect to the unique indicator variances (δ), which is known as strict factor invariance (SFI)

The MGCFA model in equation (3) is perhaps the most common approach for assessing the various types of FI (Meredith, 1993).

$$x_g = \tau_g + \Lambda_g\xi + \delta_g \tag{3}$$

The population parameters in equation (3) correspond to those in equation (1), except that they are group specific, as denoted by the g subscript. In other words, the MGCFA model allows the intercepts (τ_g), loadings (Λ_g), and unique variances (δ_g) to vary by group. The indicator covariance matrix in (2) can also be group specific, as appears in equation (4):

$$\Sigma_g = \Lambda_g\Psi_g\Lambda_g' + \Theta_g \tag{4}$$

MGCFA is used to test for the presence of various types of FI through the placement of group equality constraints on the model parameters in equations (3) and (4). For example, to assess CI, a factor model is fit to both groups where the number of latent variables and the indicators associated with them are the same between groups, but no group equality constraints are placed on any of the model parameters. If this model fits the data well, based on indices such as the comparative fit index (CFI), the root mean squared error of approximation (RMSEA), or the Standardized Root Mean Residual (SRMR), we can conclude that CI is present (Millsap, 2011). Next, we can assess MI by constraining the groups' factor loadings (Λ_g) to be equivalent, and comparing the fit of this model to the fit of the unconstrained model (3). Because the constrained and unconstrained models are nested, it is possible to compare their fit using the difference in their model χ^2values, which is itself distributed as a χ^2, with degrees of freedom equal to the difference in degrees of freedom for the two models. A statistically significant result for this χ^2_Δ statistic leads to rejection of the null hypothesis that the groups have equal loadings in the population; i.e., MI is not present in the population. On the other hand, if the χ^2_Δ is not statistically significant (i.e., MI holds), the researcher can proceed to test other types of FI, such SI and SFI.

Although it is intuitive and convenient to use, prior research has found that the χ^2_Δ test does not always maintain the nominal Type I error rate that is assumed when it is being used. For example, Chen (2007) reported that this test is very sensitive to both sample size, and to a lack of normality of the indicators. Furthermore, if the factor model is misspecified in some fashion, χ^2_Δ is not actually distributed as a chi-square statistic, and therefore yields inflated Type I error rates when used to test the various aspects of FI (Yuan & Bentler, 2004; Yuan & Chan, 2016). Other researchers have reported that for normally distributed indicator variables and sample sizes less than 500, χ^2_Δ can control Type I error at the nominal level (French & Finch, 2006). One set of alternatives that has been proposed for assessing FI involves the examination of differences in fit statistics such as the CFI for the constrained and unconstrained models. These approaches rely on the use of cut-values, such that when the difference is greater than some predetermined threshold (e.g., 0.01 or 0.005), invariance is said not to hold (Chen, 2007). Another alternative that has been recently proposed in the literature involves the use of effect sizes in the context of equivalence testing

in order to identify when FI is not present (Yuan & Chan, 2016). While certainly useful for descriptive purposes, these approaches do not provide a formal test of the null hypothesis that invariance is present. Given these mixed results for χ^2_Δ, and the lack of a true hypothesis test for assessing FI, there remains room in the invariance testing literature for an alternative approach. As is discussed below, one such alternative involves the use of the PLE approach.

Penalized Likelihood Estimator

In recent years, a number of authors have described approaches for fitting factor models, and structural equation models (SEMs) that combine elements of CFA, in which strong constraints are placed upon model parameters, with exploratory approaches that allow free estimation of certain parameters. For example, Asparouhov and Muthèn (2009) introduced exploratory structural equation models, which combines the use of an exploratory approach for defining latent variables, in conjunction with a confirmatory fitting of structural aspects of a model. A primary advantage of this modeling approach is that it does not require the imposition of strong constraints on all model parameters, thereby allowing for an exploration of factor structure, when it is not clear a priori what it should be. Other authors have approached this problem of uncertainty with regard to certain model parameters through the lens of penalized parameter estimation (e.g., Jacobucci, Grimm, & McArdle, 2016; Hirose & Yamamoto, 2015; Tutz & Schauberger, 2015). These approaches treat the estimation of model parameters in the context of CFA and SEM as a problem of sparsity, whereby some model parameters are freely estimated, and others are penalized using one of several popular approaches such as the Lasso (Tibshirani, 1996).

Huang, Chen, and Weng (2017) discussed a penalized likelihood estimator (PLE) for fitting SEMs in the context of exploratory SEM (ESEM). This approach builds upon the standard maximum likelihood estimator (MLE), by adding a penalty term to the fitting function. The standard MLE fitting function for SEM takes the form:

$$L(\theta) = -0.5ln|\Sigma(\theta)| - 0.5\big(y - \mu(\theta)\big)^T \Sigma(\theta)^{-1}\big(y - \mu(\theta)\big) \quad (5)$$

where θ =Set of parameter estimates
$\Sigma(\theta)$ =Model implied covariance matrix
$\mu(\theta)$ =Model implied mean vector
y =Observed vector.

The PLE of Huang, et al. is expressed as:

$$U(\theta,\gamma) = L(\theta) - R(\theta,\gamma) \quad (6)$$

where $R(\theta,\gamma) = \sum_{j=1}^{J} c_j \rho\big(|\theta_j|,\gamma\big)$
$\rho\big(|\theta_j|,\gamma\big)$ =Penalty function
γ =Regularization parameter
c_q =Penalty indicator for observed variable j.

When the indicator c_j is set to 0, the model parameter(s) for observed indicator j, θ_j, are freely estimated, whereas when $c_j = 1$, parameter θ_j is penalized in its estimation. The γ parameter determines the level of model complexity (i.e., penalization), where larger values of γ lead to a more sparse model; i.e., parameters with $c_j = 1$ tend toward 0. In practice, optimal values of γ are selected using comparative model fit statistics, such as the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), or some variant of those.

Huang et al. (2017) discuss the application of several approaches for fitting the PLE, including the Lasso (L1) and the Mimimax Concave Penalty (MCP). We will first describe the MCP function, of which the Lasso is a special case. The MCP function is

$$\rho MCP(t,\gamma) = \begin{cases} \gamma t - \frac{t^2}{2\delta} & if\ t \le \delta\gamma \\ 0.5\gamma^2\delta & if\ \delta\gamma < t \end{cases} \quad (7)$$

where δ =Parameter that controls the convexity of the penalty function.

Smaller values of δ make the PLE behave similarly to best subsets regression selection, whereas when $\delta \to \infty$, it becomes the Lasso estimator. Huang, et al., note that when γ is small, PLE behaves similarly to ESEM, such that many of the model parameters may take non-0 values. Conversely, when γ is large (meaning the penalty is more severe), PLE behaves much like traditional CFA/SEM, with many parameters taking the value of 0. As noted above, in practice researchers may fit several models with differing values of γ and δ, and then select the one that yields optimal fit, based on information index values.

As an example of applying the PLE in practice, a researcher may wish to apply a sparsity penalty to factor loadings linking specific indicator variables to a latent variable. In such a case, the researcher may be unsure a priori whether the indicators are in fact linked to the factor in question, given a lack of theory and/or prior empirical evidence. Through application of a sparsity penalty, such as the Lasso, estimation of these loadings can be done in an exploratory fashion so that hard constraints (e.g., loadings linking the indicators to a factor are set to 0) are not imposed on the model. In this regard, the application of PLE is very similar to that of ESEM, as described by Asparouhov and Muthèn (2009).

Invariance Assessment with Penalized Likelihood

Although the primary focus of work heretofore has been with respect to applications of PLE to situations where ESEM would be appropriate, it can also be applied to the problem of parameter invariance testing (Huang, 2018). Unlike the MGCFA approach to this problem, the PLE model does not constrain parameters for which invariance assessment is desired (e.g., factor loadings) to be equal between the groups. Rather, invariance (or a lack thereof) is expressed in terms of what portion of the values is associated with a reference component, versus what additional information is added by the group specific increment. When this increment for a group differs from that of other groups, we can conclude that there is a lack of invariance for the model parameter.

In order to provide a deeper explanation of how PLE can be used for invariance assessment, consider a loading relating indicator variable *j* to a factor.

$$\lambda_{jg} = \lambda_{jq} + \lambda_{jqg} \tag{8}$$

where λ_{jg} =Factor loading for group g on indicator variable *j*
λ_{jq} =Reference component of loading for indicator variable *j*; common across groups
λ_{jqg} =Group specific increment component.

Equation (8) shows that the factor loading for a specific group, g, is a function of a reference component that is common across groups, and a component that represents the specific contribution of the group itself, known as the increment. When the increment is 0, then the group specific factor loading is equivalent to the reference component, meaning that the group does not contribute anything unique to the loading. On the other hand, a non-zero value for λ_{jqg} indicates that membership in group g does contribute something unique to the relationship between observed indicator *j* and the factor. In other words, the loading for group g does differ from that of the reference component. In the multiple groups case, the likelihood function in (5) becomes

$$L(\theta) = -0.5\sum_{g=1}^{G} w_g \left[ln\left|\Sigma(\theta_g)\right| + tr\left(\Sigma(\theta_g)^{-1} S_g\right)\right] - 0.5\sum_{g=1}^{G} w_g \left(y_g - \mu(\theta_g)\right)^T \Sigma(\theta_g)^{-1}\left(y_g - \mu(\theta_g)\right) \tag{9}$$

where θ_g =Set of parameter estimates for group g
$\Sigma(\theta_g)$ =Model implied covariance matrix for group g
S_g =Observed covariance matrix for group g
$\mu(\theta_g)$ =Model implied mean vector for group g
y_g =Observed vector for group g
w_g =Proportion of total sample in group g.

In turn, the penalty function takes the following form:

$$R(\theta, \gamma) = \sum_{j=1}^{J} c_j \rho\left(\left|\theta_j\right|, \gamma\right) + \sum_{g=1}^{G}\sum_{j=1}^{J} c_{gj}\, \rho\left(\left|\theta_{gj}\right|, \gamma\right) \tag{10}$$

where $\rho\left(\left|\theta_{gj}\right|, \gamma\right)$ =Penalty function for group g
c_{gj} =Penalization indicators for group g

Huang (2018) provides an example involving factor loadings to describe how factor invariance assessment with PLE works in practice. We borrow this example here to provide a bit more insight regarding this approach. Huang's description of this example appears on pages 4 and 5 of his manuscript. To motivate this example, consider factor loadings for indicator variable *j* and groups 1 and 2. These can be written as:

λ_{jG1} =Factor loading for indicator *j* for group 1
λ_{jG2} =Factor loading for indicator *j* for group 2.

Applying equation (8), we can express these loadings as follows:

$$\lambda_{jG1} = \lambda_{jq} + \lambda_{jq1} \tag{11}$$

$\lambda_{jG2} = \lambda_{jq} + \lambda_{jq2}$
$\lambda_{jq1} = 0$ implies:

1. $\lambda_{jG1} = \lambda_{jq}$,
2. $\lambda_{jG2} = \lambda_{jq2} - \lambda_{jq1}$

In turn, implication 2 means that when $\lambda_{jq2} = 0$, the groups' factor loadings will be invariant; i.e., $\lambda_{jG2} = \lambda_{jG1} = \lambda_{jq}$.

In practice, MI assessment based on PLE involves first setting the referent indicator loading to some value across groups; e.g., 1. Next, the constraint of $\lambda_{jq1} = 0$ is set for all indicators. In this way, the reference component portion of each loading is defined. Third, PLE is applied to the non-referent indicators, accounting for the previous constraint of the group 1 loadings. The optimal model is then selected, using AIC, BIC, or one of their variants. If the penalized estimate of the group 2 increment component for indicator *j* is not equal to 0 ($\lambda_{jq2} \neq 0$), we would conclude that the loading is not invariant across the groups. If, on the other hand, $\lambda_{jq2} = 0$, then invariance for that loading is found to hold.

Huang (2018) investigated the performance of the PLE based invariance assessment procedure using a Monte Carlo simulation study. The study focused on the 2 groups, 1 factor measurement invariance condition. The factor was simulated to have 12 multivariate normal indicators, with factor loadings ranging between 0.6 and 0.8. A lack of invariance was simulated by reducing the loadings for the second group, such that they differed by 0.1 (small noninvariance), 0.2 (moderate noninvariance), or 0.3 (large noninvariance). Total sample sizes of 200, 400, 600, 800, and 1000 were simulated, with the groups being of equal size across conditions. The outcomes of interest were the mean squared error (MSE), squared bias, proportion of times when the true model was selected, the true positive, and false positive rates. MCP was the only penalty method included in the study, given that the Lasso is a special case of MCP. Results presented by Huang demonstrated that PLE for invariance testing has real promise for application in practice. More specifically, when invariance held between the groups (i.e., group loadings didn't differ), or the loadings differed in either the moderate or large range, MCP with the BIC used to select optimal settings of γ and δ performed well in terms of identifying the correct model, particularly when the sample size was large. For small samples, or when the level of invariance was small, AIC yielded the highest rate of correct model identification. Finally, Huang found that when the sample size was small and invariance held, neither AIC nor BIC were able to identify the correct model very well.

Goals of the Current Study

The primary goal of the current simulation study is to extend upon the work of Huang (2018) by comparing the performance of PLE for invariance testing with that of the standard MLE based approach. Huang demonstrated that the PLE approach holds promise for researchers in practice, but more simulations are needed to further understand how well it performs across a variety of conditions, and vis-à-vis the standard MGCFA approach based on MLE. Thus, as is described below, this study extends several of the conditions examined by Huang, and includes others that were not investigated in the earlier research. The research questions and hypotheses to be addressed in this study are:

1. How does the Type I error rate of PLE invariance testing compare to that of MLE? It is hypothesized that the Type I error rate of PLE will be lower than that of MLE, and that it will be at or below the nominal 0.05 level.

2. How does the power of PLE invariance testing compare to that of MLE? It is hypothesized that power for PLE will be lower than that of MLE for small sample sizes, but comparable for larger samples.

3. How does parameter estimation bias of PLE compare to that of MLE? It is hypothesized that bias in the PLE based estimates will be greater than that of MLE, which will exhibit little or no such bias.

4. How do coverage rates of PLE compare to those of MLE? It is hypothesized that coverage rates for the two approaches will be comparable.

Method

In order to address the research questions and hypotheses outlined above, a Monte Carlo simulation study was used, with 1000 replications per combination of the conditions described below. When replications did not converge, additional simulations were run in order to obtain the requisite number of 1000 per combination of study conditions. Data were generated using Mplus version 8.0 (Muthèn & Muthèn, 2017), and were analyzed using the `lavaan` (Rossel, 2012) and `lslx` (Huang, 2018) libraries in the R software package, version 3.4.3 (R Development Core Team, 2016). Data were generated for 2 groups, with 1 factor, and measurement invariance was assessed using the χ^2_Δ statistic in conjunction with MLE, as well as using the PLE based approach outlined above. The indicator variables were generated from the multivariate standard normal distribution, with factor model intercepts of 0 for all indicators, and variances for all factors simulated to be 1. Factor loadings for group 1 were simulated to be 1, and were varied for group 2 in order to induce a lack of invariance, as is described below. When factor loadings differed, such differences were simulated for 10% of the indicator variables. All other factor model parameters were simulated to be invariant across groups. The following conditions were manipulated in the study.

Sample size

Given that sample size has been shown to have an impact on the performance of both the PLE invariance procedure and the χ^2_Δ statistic (Chen, 2007; Huang, 2018), a variety of sample sizes were simulated in the current study. The total sample size conditions included were 50, 100, 200, 400, 600, 800, and 1000. The groups were simulated to have equal sample sizes, leading to individual group sizes of 25, 50, 100, 200, 300, 400, and 500. These overall values were designed to reflect cases from very small (50) to large (1000).

Number of Indicator Variables

Data were simulated to have either 10, 20, or 30 observed indicator variables. These values were selected to represent a variety of real world cases, from a fairly small number (10) to a relatively large number of indicators (30).

Magnitude of Group Factor Loading Difference

In order to assess and compare both the Type I error rate coverage, and the power of the MLE and PLE approaches, a variety of group factor loading differences were simulated. In the Type I error rate case, the loadings were simulated to be equivalent between the groups. In order to assess power, group loading differences were simulated to be 0.1, 0.2, 0.3, 0.4, and 0.5, in order to represent a range of noninvariance from small (0.1) to large (0.5). Measurement noninvariance was simulated by subtracting the group loading difference from the group 1 loading, in order to obtain the group 2 loading. For example, in the 10 indicators 0.2 loading difference magnitude condition, the loadings for groups 1 and 2 were simulated as below:

Group 1 loadings: 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
Group 2 loadings: 1, 0.8, 1, 1, 1, 1, 1, 1, 1, 1

Estimation Method

Three parameter estimation methods were used in this study: MLE, Lasso, and MCP. As noted above, the Lasso is a special case of MCP where $\delta = \infty$. The Lasso is included in the study in order to ascertain whether this simpler PLE approach works comparably to, or better than, the more complex MCP. The MCP is more complex to use because it requires finding the optimal settings for two penalty parameters, δ and γ, whereas the Lasso only requires finding the optimal setting for γ. MLE is included in the study because it is generally regarded as the standard estimation technique to use with normally distributed indicator variables.

With respect to the invariance testing, the χ^2_Δ statistic was used to test the null hypothesis that MI held for the two groups. In terms of assessing MI using PLE, the steps outlined in Huang (2018), and which were described in some detail above, were employed in this study. To summarize, the referent indicator (variable 1) was constrained to be equal to a set value, such as 1, for both groups. For the remaining indicators, the group 1 specific increment was set to 0 ($\lambda_{jG1} = 0$) in order to define the reference component for each loading. PLE was then applied to these non-referent indicators, accounting for the fact that the group 1 loadings were constrained as described above. The optimal model was selected using an information index, and if the penalized estimate of the group 2 specific increment component for indicator j was not equal to 0 for this optimal model, the loading was not assumed to be invariant across the groups; i.e., MI did not hold. Conversely, if the PLE group 2 specific increment was 0, then invariance for that loading was assumed. The AIC, BIC, and CAIC were each used in the current study, with the CAIC consistently providing optimal results. Therefore, results using only the CAIC are reported below.

Study Outcomes of Interest

Several outcome variables were of interest in this study, including the Type I error rate and power for identifying a lack of invariance, relative parameter estimation bias for the target group 2 loading, and coverage rates for the target group 2 loading. In this study, the group 2 factor loading estimate for indicator variable 2 served as the target. For each estimator (MLE, Lasso, MCP) the Type I error rate was simply calculated as the proportion of replications for which factor loadings were identified as different between the groups when they were simulated to be equal between groups. Likewise, power was the proportion of replications for which the factor loadings were identified to be different between the groups, when in fact they were simulated to differ between the groups. For a given replication, the relative bias for the target group 2 factor loading was calculated as follows:

$$RB = \frac{\lambda_j - \widehat{\lambda_j}}{\lambda_j} \tag{12}$$

where λ_j =Population factor loading

$\widehat{\lambda_j}$ =Estimated factor loading.

The mean of these values was then taken across the replications for a specific combination of study conditions. Parameter coverage was the proportion of replications for which the population value for the target group 2 loading fell within the confidence interval. Values for each of these outcomes were estimated for each estimation method.

In order to identify which of the manipulated variables, and their interactions were important in terms of the study outcome variables, analysis of variance (ANOVA) was used, in conjunction with the η^2 effect size. Terms of the ANOVA model that were statistically significant were identified as important with respect to the outcomes. In addition to statistical significance, the effect size for the effect is also reported in the results below. Separate ANOVA models were used for each of the outcome variables. For the Type I error and power results, rates were summarized across replications for each combination of conditions prior to the application of the ANOVA.

Results

Type I Error Rate

The ANOVA for Type I error rate identified the interaction of sample size by estimation method as the only statistically significant term in the model ($F_{12,22}$ = 2.349, p = 0.008, η^2 = 0.668). Figure 1 displays the Type I error rate by sample size and estimation method. A reference line has been placed at the nominal 0.05 error rate. These results show that, as has been reported in prior literature (Chen, 2007; Yuan & Bentler, 2004), the Type I error rate for the standard Chi-square difference test based on MLE is somewhat inflated. The greatest such inflation occurred for samples of 50 and 100 (25 and 50 per group). At the other end of the spectrum, the Type I error rate for the Lasso estimator was well below the nominal 0.05 level, with the highest value be approximately 0.016 for a sample size of 50. In all other sample size conditions, the error rate was below 0.01. The Type I error rate for the MCP estimator was above the nominal 0.05 level across sample size conditions, generally staying close to 0.06, and was below that of the MLE approach in all cases. Finally, based on guidelines laid out by Bradley (1978), only the error rate of MCP can be classified as in control across all sample size conditions, as it consistently lay between 0.025 and 0.075.

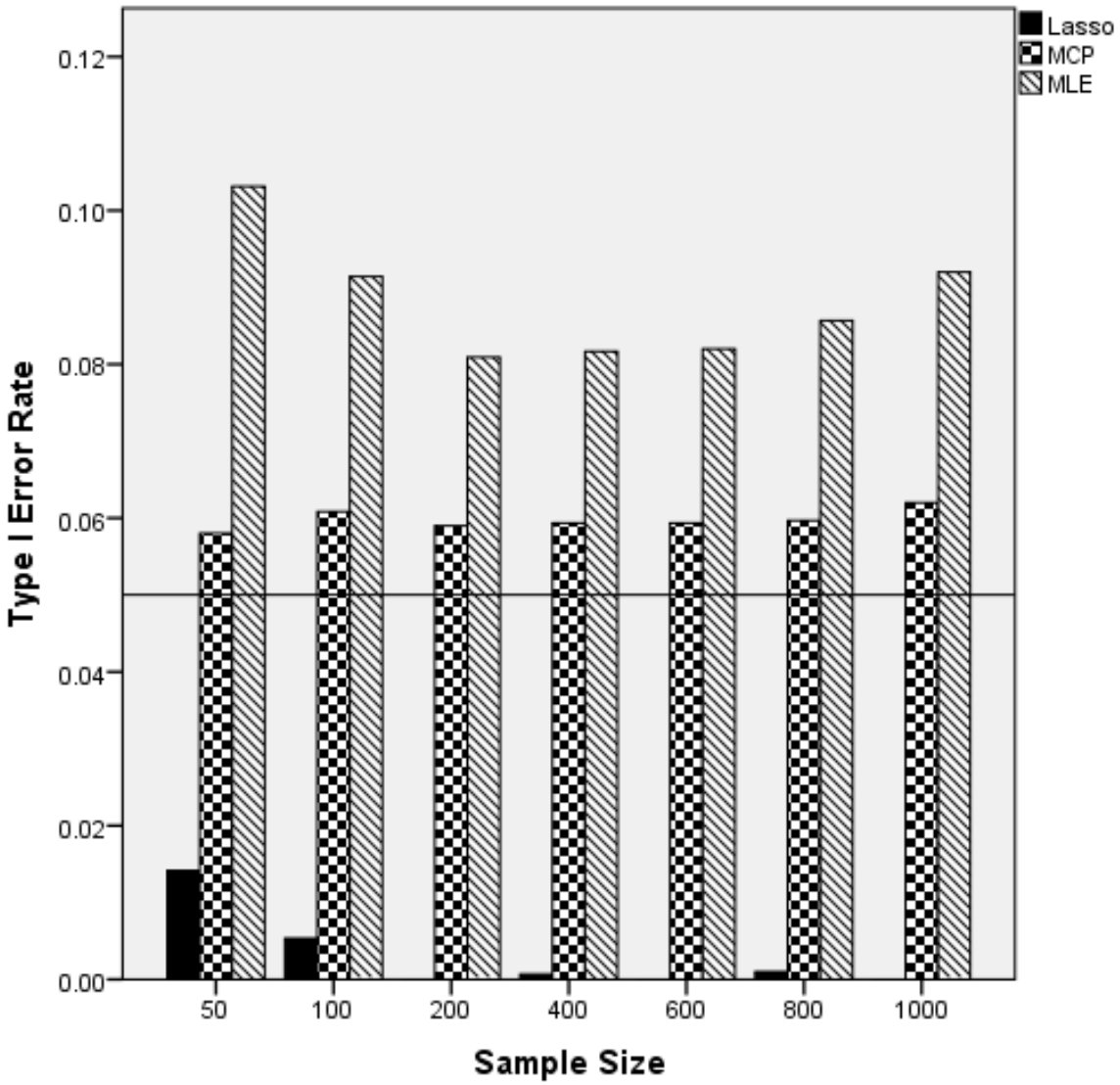

Figure 1. Type I error rate by sample size and estimation method.
Note. Reference line at nominal Type I error rate of 0.05.

Power

ANOVA identified the interactions of estimation method by sample size by magnitude of group loading difference ($F_{48,146} = 4.215, p < 0.001, \eta^2 0.581$), and estimation method by number of indicators ($F_{4,148} = 9.781, p < 0.001, \eta^2 0.211$) as being statistically significantly related to power for detecting measurement noninvariance. Figure 2 includes power rates by estimation method, sample size, and magnitude of group loading difference. When loadings differed by 0.4 or 0.5, and samples were 800 or 1000, power for the Lasso was comparable to that of the other two methods. In all other cases, the Lasso estimator yielded lower power than did MCP or MLE. Across all conditions, MCP had comparable or higher power for detecting group loading differences than did MLE. More specifically, when group loadings differed by 0.2 or 0.1, MCP had higher power than MLE, except for a sample size of 50 (25 per group). When the loadings differed by 0.4 or 0.5, the power rates of MCP and MLE were comparable for samples of 400 or more. For samples of 200 or fewer and loading differences of 0.4 or higher, MCP had higher power than did MLE. For a group loading difference of 0.3, MCP exhibited higher power rates across sample sizes, except when *N*=50. Finally, when interpreting these power results, it is important to note that under many conditions, the differences in power rates among the methods, particularly MCP and MLE, were fairly small. Thus, although MCP did often have slightly higher power rates than those of MLE, the differences were not very large.

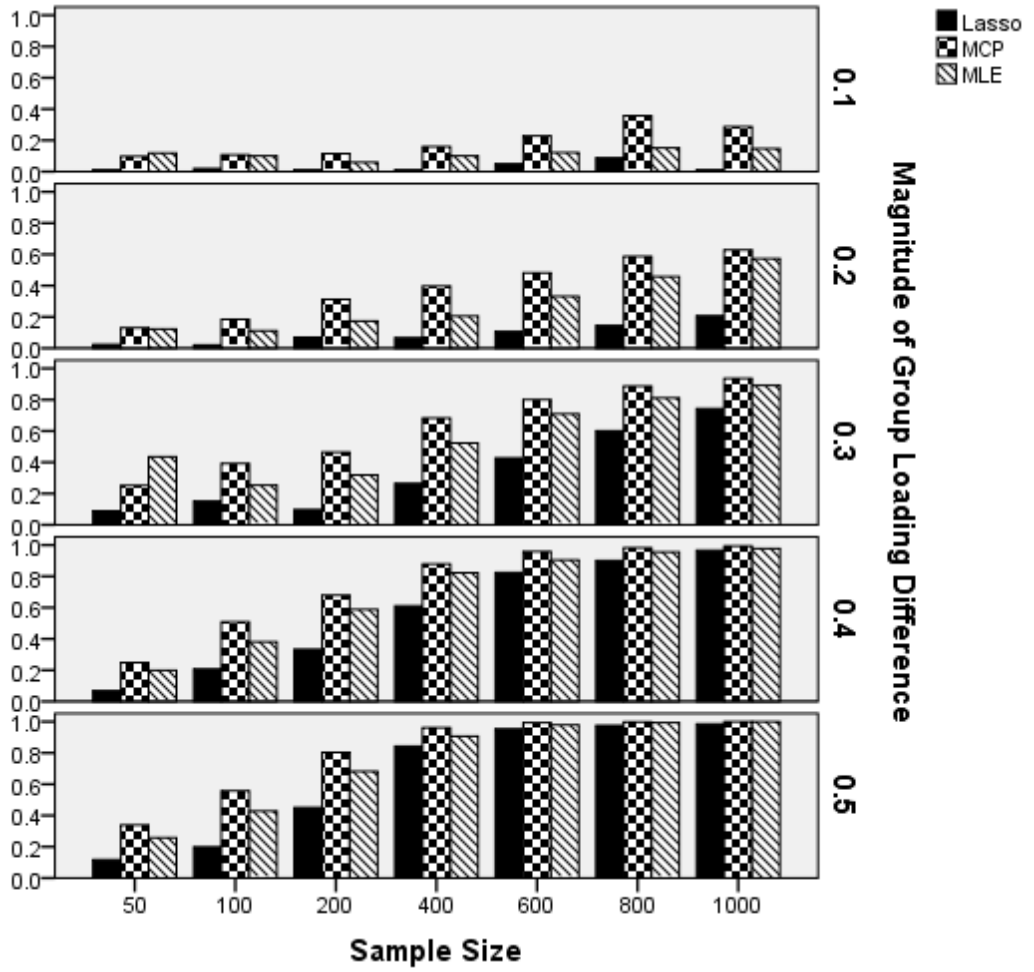

Figure 2. Power rates by estimation method, sample size, and magnitude of group loading difference.

Figure 3 displays power rates by estimation method and number of indicator variables. This graph reinforces the finding in Figure 2 that MCP generally displayed higher power rates for detecting noninvariance of factor loadings than did MLE, and that Lasso had the lowest power. However, it is also important to note that in many of these cases, power rates for the MCP and MLE, in particular, were very close to one another. Thus, although MCP did exhibit higher power rates, the differences were often not very large. In addition, both MCP and MLE yielded higher power for more indicator variables. This higher power associated with a greater number of indicators was more pronounced for MLE than for MCP, such that for 30 indicators the power rates for the two methods were closer in magnitude than was the case for 10 or 20 indicator variables.

Parameter Estimation Bias

The results of the ANOVA revealed that the interactions of estimation method by sample size, and magnitude of group loading difference ($F_{60,168} = 3.419, p < 0.001, \eta^2 0.550$), and estimation method by number of indicator variables by magnitude of group loading difference ($F_{20,168} = 1.860, p < 0.001, \eta^2 0.181$) were statistically significant. Figure 4 displays the relative parameter estimation bias for the group 2 factor loadings that were simulated to be noninvariant by the estimation method, magnitude of the group loading difference, and sample size. These results demonstrate that across conditions, the relative bias of the MLE derived loadings was comparable to, or in most cases lower, than that of either MCP or Lasso. In addition, for group loading differences of 0.2 or less, and sample sizes of 50 or 100, the MCP estimator yielded estimates with higher relative bias than did the Lasso estimator. For larger group loading differences, the Lasso estimator exhibited greater bias than did MCP at these sample size levels. Finally, bias declined for MCP and Lasso concomitantly with increases in the sample size.

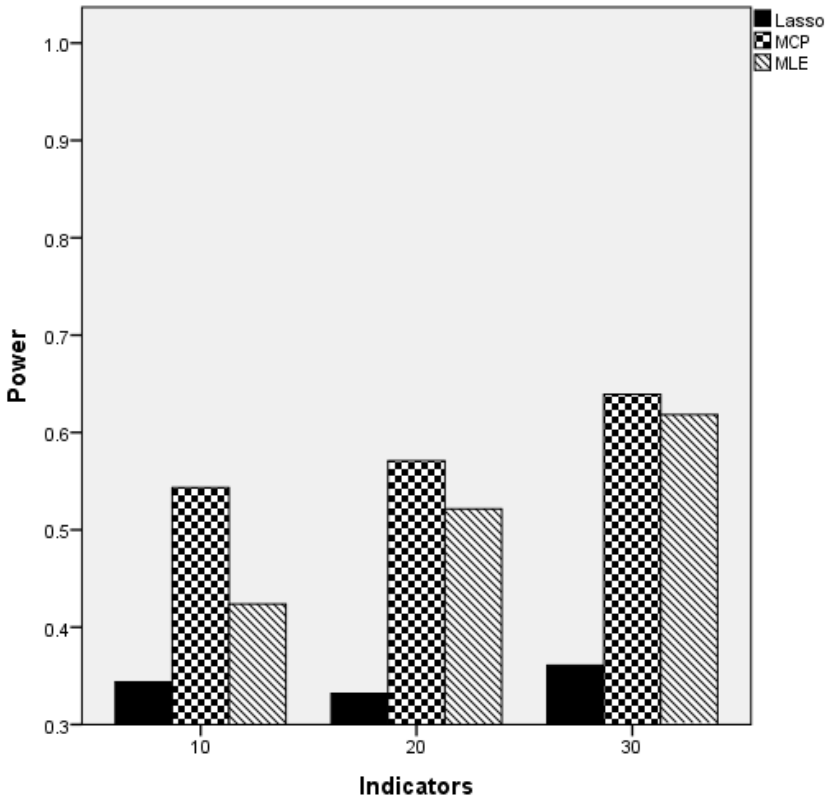

Figure 3. Power rates by estimation method and number of indicators.

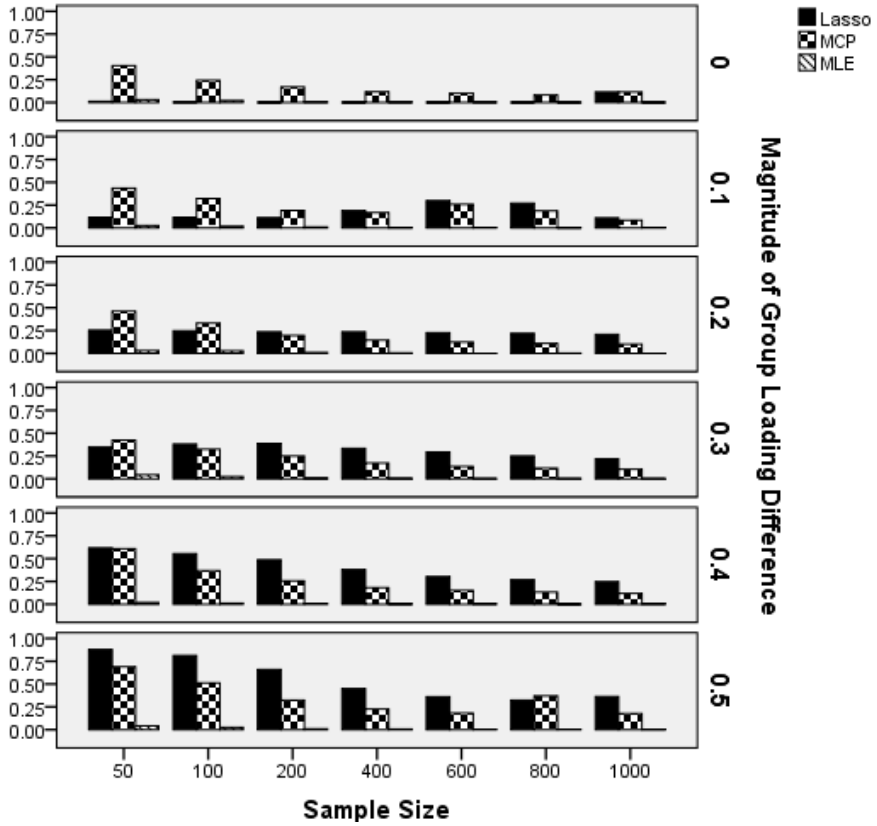

Figure 4. Relative parameter estimation bias for group 2 noninvariant factor loadings by estimation method, sample size, and magnitude of group loading difference.

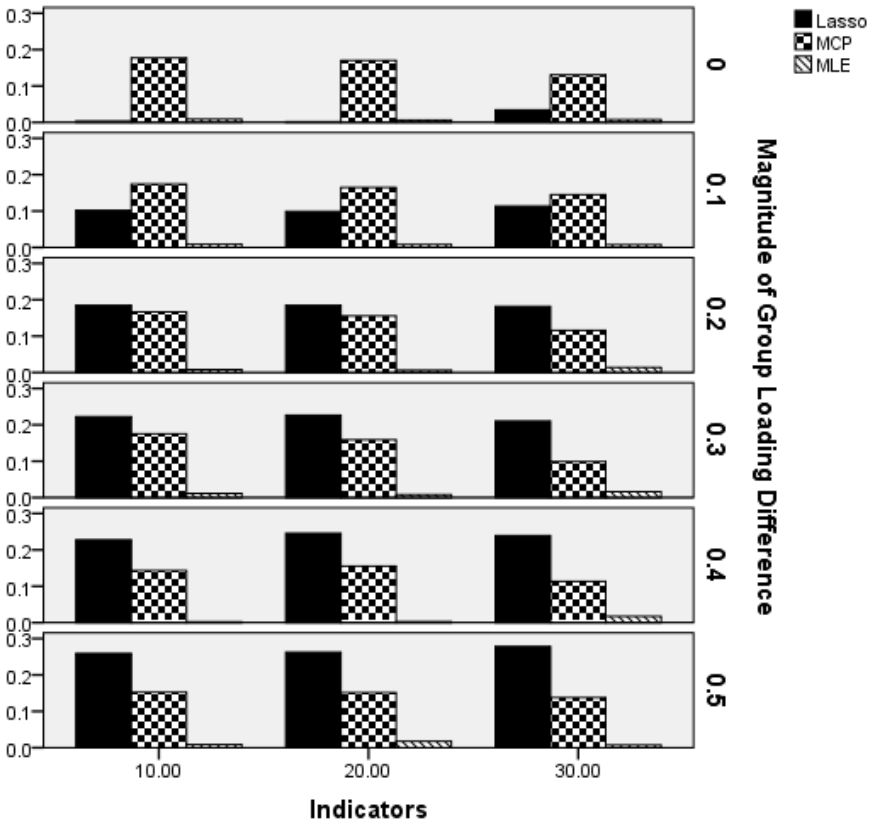

Figure 5. Relative parameter estimation bias for group 2 noninvariant factor loadings by estimation method, number of indicators, and magnitude of group loading difference.

Figure 5 displays the relative bias in the group 2 noninvariant factor loadings by the estimation method, number of indicators, and the magnitude of the group loading difference. As was evident in Figure 4, relative bias was lowest for the MLE loadings across conditions. The relative bias was highest for MCP when the magnitude of group loading difference was 0 or 0.1, above which the Lasso estimator exhibited the largest relative bias. An examination of Figure 5 reveals that the relative bias in MCP was stable across the magnitude of group loading differences, but that of the Lasso estimator increased concomitantly with increases in the magnitude of the group loading difference.

Coverage Rates

Based upon the results of the ANOVA, the interaction of estimation method by magnitude of the group loading difference by the sample size ($F_{60,168} = 3.389, p < 0.001, \eta^2 0.548$) was the only term that was statistically significantly associated with the coverage rates for the group 2 noninvariant parameter loadings. The coverage rates by estimation method, sample size, and magnitude of group loading difference appears in Figure 6. A reference line appears at the nominal 0.95 level. The coverage rates for the MCP and MLE approaches were at the nominal level across all conditions. On the other hand, the coverage rates for the Lasso estimator were below the nominal 0.95 level when the group factor loadings differed by 0.3, and the sample size was 600 or more, as well as when the loading difference was 0.4 or 0.5, and the sample size was 200 or more.

Discussion

The purpose of the current study was to investigate the performance of an approach for measurement invariance testing based on the work of Huang et al. (2017) and Huang (2018), using a PLE to fit factor models. As has been noted in the literature (Yuan & Bentler, 2004; Yuan & Chan, 2016) the standard MGCFA approach based on the Chi-square difference test has been shown to yield elevated Type I error rates. Thus, the goal of the current work was to extend prior work by Huang to ascertain whether the PLE approach might yield a viable alternative to the standard χ^2_Δ test statistic. As noted above, despite continuing concerns regarding the potential for Type I error inflation, practicing researchers continue to use the χ^2_Δ test as a primary way to investigate questions around factor invariance.

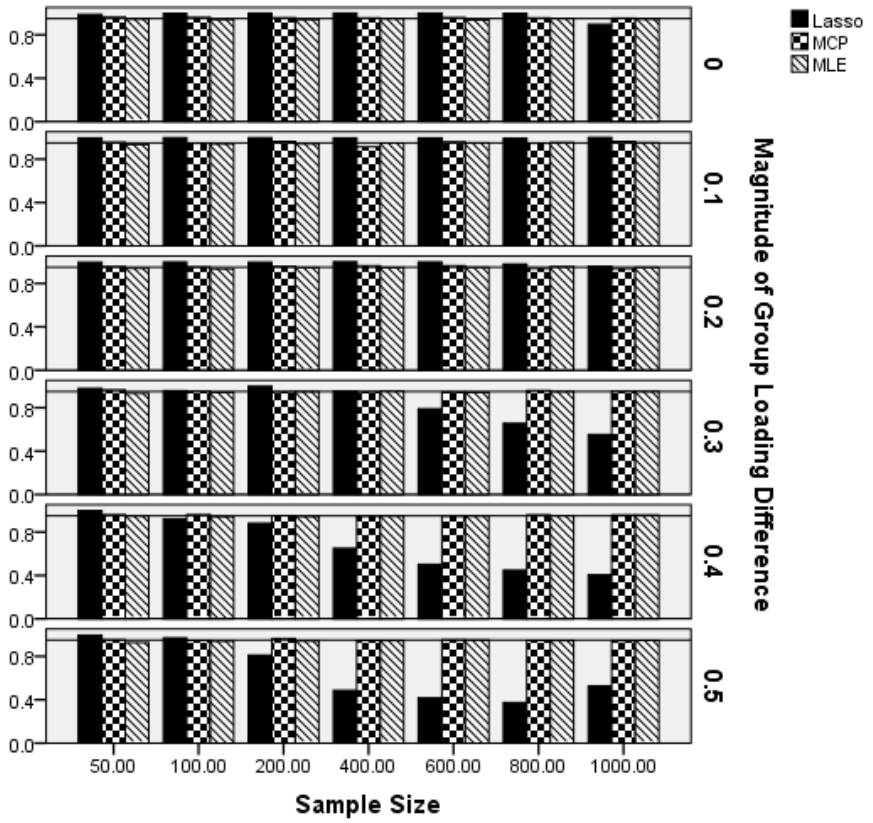

Figure 6. Coverage rates for group 2 noninvariant factor loadings by estimation method, sample size, and magnitude of group loading difference.

Note. Reference line at the nominal 0.95 level.

Alternatives for invariance assessment, such as equivalence testing based upon model fit statistics, while promising, do not provide the researcher with an explicit test of the null hypothesis that factor model parameters are invariant across groups.

Thus, it would appear to be useful for a reliable such test to be identified for use in practice.

This study was framed around 4 research questions and associated hypotheses. The first research hypothesis stated that the Type I error rate for the PLE techniques would be lower than that of the χ^2_Δtest based on MLE. The results presented above support this hypothesis. The two PLE based approaches did have lower Type I error rates than did the MLE, with the Lasso having the lowest values. MCP yielded error rates that were consistently around 0.06, which is considered in control based upon recommendations by Bradley (1978). Such cannot be said for either the Lasso or MLE results. The second hypothesis asserted that power would be lower for the PLE based approaches than for MLE. This hypothesis was supported with respect to the Lasso, which did yield lower power results than did MLE. However, the MCP invariance test yielded higher power than that of MLE, except for the larger sample size conditions, in which case the two methods had similar power rates. The third hypothesis was that MLE would have little to no parameter estimation bias, whereas the PLE based approaches would exhibit greater such bias. This hypothesis was supported by the results presented above. The final hypothesis was that coverage rates for the MLE and PLE based methods would be comparable to one another. The results presented above partially support this hypothesis. The coverage rates for MLE and MCP were indeed comparable, and at the 0.95 level across study conditions. however, the coverage rates for the Lasso estimator were below the nominal level for larger samples and a greater degree of group loading difference.

The results of this study have several implications for researchers and practitioners. First, under conditions similar to those simulated in this study, MCP for invariance testing would appear to present a viable alternative to the standard χ^2_Δ test based on MLE. Given that its Type I error rate was found to be in control, and its power was comparable to that of χ^2_Δ, researchers should certainly consider using it when conducting invariance tests. On the other hand, the Lasso estimator based approach does not seem as

promising for this purpose. While it did consistently yield Type I error rates that were well below the nominal 0.05 level, its power rates were also lower than those of the other two methods, in some cases quite a bit lower. It is possible that this lower power is a function of the higher degree of bias that was evident in the estimates produced by the Lasso. The results presented above demonstrated that this bias was larger when the difference between the group factor loadings was larger. Recall from the methods section that group differences were simulated by reducing the noninvariant loadings for group 2 by the amount of group difference. Both PLE techniques examined here work by reducing the magnitude of the penalized parameters. In the context of invariance testing, one of these parameters is the effect of loadings of the second group. Thus, when the shrinkage penalty is applied to this parameter, it would be anticipated that the estimate would exhibit more bias than would MLE, which is precisely what was found here.

Directions for Future Research

The current study was designed to extend work by Huang (2018) assessing the PLE approach for testing measurement invariance. As such, a number of simulation conditions were selected so as to provide insights into its performance across various sample size, number of indicators, and group loading difference magnitude conditions. As with all studies, there were conditions that were not examined in the current work, but which need to be investigated in future studies. For example, a wider set of conditions for the proportion of noninvariant loadings should be researched. In this study, 0% and 10% of loadings were simulated to differ between the groups. Future work should investigate higher proportions of noninvariant loadings, in order to provide insights into the performance of MCP and Lasso in such cases. In addition, future research should investigate the performance of the PLE approaches with respect to scalar invariance assessment. MCP, in particular, showed promise for testing the null hypothesis of no measurement invariance, but this cannot be seen as evidence that it will perform similarly when applied to the assessment of factor intercept differences between groups. Finally, future work in this area needs to examine how the PLE methods for assessing factor invariance perform with categorical indicator variables. This is particularly important given that in many applications in the social sciences invariance assessment is applied to scales consisting of dichotomous or polytmous items.

Conclusions

Taken together with the promising findings from Huang (2018), and Huang et al. (2017), the results of the current study provide support for the use of PLE for invariance testing, in the form of the MCP estimator. Certainly, more work investigating its performance under a wider variety of conditions is needed. However, the current results provide evidence supporting the use of PLE for invariance assessment. The MCP estimator exhibited superior control of the Type I error rate to that of MLE, and had higher or comparable power across conditions. This approach would seem to be quite useful when used in conjunction with an effect size based approach for invariance assessment, such as the equivalence procedure of Yuan, Chan, Marcoulides, and Bentler (2016). In addition, though the relative bias for MCP was higher than that of MLE, it was consistently around 0.17. This would mean that for a population factor loading of 0.8 the MCP estimate might be approximately 0.64. Certainly these values differ, but in both cases the fact that there exists a relationship between the indicator and the factor is clear. Therefore, it could be argued that the level of bias in the MCP estimate would not alter the final conclusions drawn regarding the nature of the factor models. In addition, researchers could use the MCP approach to test for factor invariance, but then rely on consulting both the MCP and MLE factor loading estimates in order to gain insights into the factor structure itself. In summary, the results of this study support the utility of the PLE approach, in the form of MCP, for assessing measurement invariance.

References

Asparouhov, T., & Muthèn, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling, 16,* 397-438.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*(2), 144-152.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 465-504.

Dorans, N. J., & Cook, L. L. (Eds.). (2016). *Fairness in educational assessment and measurement.* New York: Routledge.

French, B. F., & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling, 13,* 378-402.

Hirose, K., & Yamamoto, M. (2015). Sparse estimation via nonconcave penalized likelihood in factor analysis model. *Statistical Computing, 25,* 863-875.

Huang, P.-H. (2018). A penalized likelihood method for multi-group structural equation modeling. *British Journal of Mathematical and Statistical Psychology, 71*(3), 499-522.

Huang, P. H., Chen, H., & Weng, L. J. (2017). A penalized likelihood method for structural equation modeling. *Psychometrika, 82*(2), 329–354.

Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized structural equation modeling. *Structural Equation Modeling, 23*(4), 555-566.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika, 58*(4), 525-543.

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

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (7th ed.). Los Angeles: Authors.

R Development Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rossel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1-36.

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

Tutz, G., & Schauberger, G. (2015). A penalty approach to differential item functioning in Rasch models. *Psychometrika 80*(1), 21–43.

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMMS data. *Practical Assessment, Research & Evaluation, 12*(3), 1-26.

Yuan, K.-H., & Bentler, P. M. (2004). On chi-square difference and Z tests in mean and covariance structure analysis when the base model is misspecified. *Educational and Psychological Measurement, 64*, 737-757.

Yuan, K.-H., & Chan, W. (2016). Measurement invariance via multigroup SEM: Issues and solutions with chi-square difference tests. *Psychological Methods, 21*(3), 405-426.

Yuan, K.-H, Chan, W., Marcoulides, G. A., & Bentler, P. M. (2016). Assessing structural equation models by equivalence testing with adjusted fit indexes. *Structural Equation Modeling, 23*(3), 319-330.

Send correspondence to:
W. Holmes Finch
Ball State University
Email: whfinch@bsu.edu