Routledge
Taylor & Francis Group

Structural Equation Modeling: A Multidisciplinary Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hsem20

Regularized Structural Equation Modeling to Detect Measurement Bias: Evaluation of Lasso, Adaptive Lasso, and Elastic Net

Xinya Liang & Ross Jacobucci

To cite this article: Xinya Liang & Ross Jacobucci (2020) Regularized Structural Equation Modeling to Detect Measurement Bias: Evaluation of Lasso, Adaptive Lasso, and Elastic Net, Structural Equation Modeling: A Multidisciplinary Journal, 27:5, 722-734, DOI: 10.1080/10705511.2019.1693273

To link to this article: https://doi.org/10.1080/10705511.2019.1693273

STRUCTURAL EQUATION MODELING: A MULTIDISCIPLINARY JOURNAL
2020, VOL. 27, NO. 5, 722–734
https://doi.org/10.1080/10705511.2019.1693273

Routledge
Taylor & Francis Group

Regularized Structural Equation Modeling to Detect Measurement Bias: Evaluation of Lasso, Adaptive Lasso, and Elastic Net

Xinya Liang [1] and Ross Jacobucci[2]

[1]University of Arkansas; [2]University of Notre Dame

ABSTRACT

Correct detection of measurement bias could help researchers revise models or refine psychological scales. Measurement bias detection can be viewed as a variable-selection problem, in which biased items are optimally selected from a set of items. This study investigated a number of regularization methods: ridge, lasso, elastic net (enet) and adaptive lasso (alasso), in comparison with maximum likelihood estimation (MLE) for detecting various forms of measurement bias in regard to a continuous violator using restricted factor analysis. Particularly, complex structural equation models with relatively small sample sizes were the study focus. Through a simulation study and an empirical example, results indicated that the enet outperformed other methods in small samples for identifying biased items. The alasso yielded low false positive rates for non-biased items outside of a high number of biased items. MLE performed well for the overall estimation of biased items.

KEYWORDS

Adaptive lasso; elastic net; measurement bias detection; regularized structural equation modeling

Detecting measurement bias has been a long-standing issue in measurement invariance literature (Meredith, 1993; Millsap, 2011). Measurement bias can be examined in regard to any variables. In the presence of measurement bias, the differences in observed scores may not solely reflect the differences in the traits or construct levels, but depend on the group membership or the variable affecting the measurement. Locating biased items can be challenging using conventional approaches (Millsap & Everson, 1993; Oort, 1992), mainly due to the use of iterative model comparisons (Millsap & Kwok, 2004) and the difficulty in locating an unbiased anchor item (French & Finch, 2008). If the sample size is small relative to a complex model with a large number of parameters, estimation issues may arise which lead to model non-convergence or inaccurate testing results (Barrett & Kline, 1981; Deng, Yang, & Marcoulides, 2018).

Measurement bias detection can be viewed as a variable-selection problem, in which the aim is to optimally select biased items from a set of test items. Current literature has proposed a rich variety of variable-selection methods, with one prominent method being the *regularized regression* (e.g., Hastie, Tibshirani, & Friedman, 2009). The use of regularization is common in sparse modeling where only a few parameters are non-zero in the population. Regularization methods attempt to shrink the estimates of *noise* parameters (nearly-zero effects that represent the random chance) to zero, while maintaining satisfactory estimates of *signal* parameters (non-trivial effects in the population). Regularization is formulated through imposing a penalty term on the model fit function. With an appropriate choice of penalty, regularization is able to perform simultaneous variable selection (e.g., select-biased items) and parameter estimation, and achieve a balance between model fit and model complexity (Hoerl & Kennard, 1970). Moreover, regularization methods have been shown to work well in scenarios with a large number of variables and small sample sizes (Fan & Peng, 2004; Jacobucci, Brandmaier, & Kievit, 2019).

Though developed in regression with only observed variables, regularization methods have recently been extended to varying fields of latent variable modeling (e.g., Huang, 2018; Huang, Chen, & Weng, 2017; Jacobucci, Grimm, & McArdle, 2016; Lindstrøm & Dahl, 2019; Serang, Jacobucci, Brimhall, & Grimm, 2017; Tutz & Schauberger, 2015). Particularly in measurement invariance studies, Huang (2018) and Lindstrøm and Dahl (2019) proposed using penalized likelihood methods to investigate factorial invariance (Meredith, 1993) in multi-group structural equation modeling (SEM), given that the heterogeneity across groups can be examined by a sparse parameter structure. In other words, the increment components of parameters (e.g., factor loadings, intercepts) from one group to another were treated as penalized parameters, and the sparsity pattern of the increment components was examined. Finch (2018) extended the work by Huang (2018) through comparing regularization methods in additional simulation conditions. All these studies showed that regularization methods were promising alternatives in measurement invariance assessments. However, previous studies only considered two-group comparisons using multi-group SEM. More investigations are needed to understand how different regularization methods perform when measurement bias occurs in regard to a continuous covariate. One common approach for handling continuous covariates is

CONTACT Xinya Liang xl014@uark.edu Educational Statistics and Research Methods, Department of Rehabilitation, Human Resources and Communication Disorders, University of Arkansas, 751 W. Maple St., Fayetteville, AR 72701.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hsem.

based on the *restricted factor analysis* (RFA; Oort, 1992). The motivation of this study was to investigate regularization methods for detecting various forms of measurement bias using RFA models, particularly in complex SEM models with relatively small sample sizes.

Measurement bias detection using RFA

Measurement bias

Measurement bias can be defined as a violation of measurement invariance, conditional on the trait level (Mellenbergh, 1989; Meredith, 1993). Measurement invariance holds when participants have the same response behaviors to the measurement instrument given the same latent traits (η) regardless of individual characteristics (V; e.g., gender, ethnicity, age, etc.), expressed as:

$$f(X|\eta, V) = f(X|\eta), \tag{1}$$

where X is observed items measuring the latent traits η, V is called a *violator* of measurement invariance, and f is a conditional probability distribution. If the conditional probability of X on η is not independent of V, that is, $f(X|\eta, V) \neq f(X|\eta)$, the measurement is biased with respect to V. If V is a categorical variable indicating the group membership, one common approach is to use multi-group SEM to investigate, for example, whether the violation of loading equality (non-uniform bias) or violation of intercept equality (uniform bias) occurs across groups. If V is on an arbitrary measurement scale (continuous or discrete) or multiple violators Vs are present, one may employ alternative approaches such as the RFA model or the multiple indicator multiple cause (MIMIC) model (Muthén, 1989). The two models differ only in modeling the relationship between V and η. In RFA, V and η correlate, comparing to the MIMIC model where V predicts η. Both models yield identical results for measurement bias detection.

The RFA model

The RFA model includes a set of covariates in the common factor analysis models, expressed as:

$$\mathbf{x}_i = \mathbf{\Lambda}\boldsymbol{\eta}_i + \mathbf{B}\mathbf{v}_i + \mathbf{\Gamma}\boldsymbol{\eta}_i\mathbf{v}_i + \boldsymbol{\delta}_i, \tag{2}$$

where for person i, $\mathbf{\Lambda}$ is the factor loading matrix defining the associations between p observed variables $\mathbf{x}$ and q latent factors $\boldsymbol{\eta}$, $\mathbf{B}$ contains regression coefficients assessing the effect of violator scores $\mathbf{v}$ on $\mathbf{x}$, $\mathbf{\Gamma}$ contains regression coefficients assessing the impact of interaction $\boldsymbol{\eta}^*\mathbf{v}$ on $\mathbf{x}$, and $\boldsymbol{\delta}$ is a vector of measurement errors. If there is a direct effect from $\mathbf{v}$ or $\boldsymbol{\eta}^*\mathbf{v}$ to $\mathbf{x}$, the measurement of $\mathbf{x}$ is biased with respect to $\mathbf{v}$; otherwise, the $\mathbf{x}$ variables are measurement invariant. Measurement bias can be classified into uniform bias and non-uniform bias (Mellenbergh, 1989). Uniform bias presents if item intercepts differ respective to $\mathbf{v}$, measured by direct paths from $\mathbf{v}$ to $\mathbf{x}$. Non-uniform bias occurs if both $\boldsymbol{\eta}$ and $\mathbf{v}$ cause the difference on $\mathbf{x}$ (i.e., non-invariant intercepts and loadings), measured by direct paths from $\boldsymbol{\eta}^*\mathbf{v}$ to $\mathbf{x}$. The non-uniform bias manifested as non-linear interaction effects is less straightforward to test, because the interactions often times do not follow multivariate normal distributions. Common methods to handle the interaction effects include product indicator approaches (Chin, Marcolin, & Newsted, 2003) and latent moderated structural equations (LMS; Klein & Moosbrugger, 2000). The LMS has been shown to provide efficient and unbiased parameter estimates.

Regularization methods

Overview

Regularization can be conducted using likelihood-based approaches or Bayesian approaches. In likelihood-based approaches, a cautiously selected penalty term can be imposed on the discrepancy function of any estimator to achieve the selection of non-trivial parameters. The increase in penalty could lead to more parameter estimates being shrunk toward zero. As small estimates can be shrunk directly to zero, regularization is typically used to reduce the model complexity/dimension and offer an easy interpretation of the model structure. With Bayesian estimation, regularization is achieved through manipulating the prior distributions on model coefficients. The priors are often chosen with small credibility intervals (e.g., normal distribution priors with zero means and small variances), such that the posterior distributions are constrained to center near zero, leading to posterior point estimates penalized toward zero. Literature in modern statistics has provided a range of shrinkage methods for high-dimensional inferences, which aim at regularizing the global noises while retaining good recovery of local signals. These shrinkage methods use the so-called global-local (G-L) priors (Polson & Scott, 2010), featuring the use of multivariate scale mixtures of normal priors in a hierarchical model (e.g., Carvalho, Polson, & Scott, 2010).

Regularization in SEM

Regularized SEM was first proposed to identify cross-loading patterns in SEM (Huang et al., 2017; Jacobucci et al., 2016), and then further developed to multi-group SEM for measurement invariance testing (Huang, 2018). In regularized SEM, a penalty term $\lambda P(\cdot)$ is added to the maximum likelihood estimation (MLE) fit function F_{MLE} as follows:

$$F_{\text{RegSEM}} = F_{\text{MLE}} + \lambda P(\cdot), \tag{3}$$

where F_{MLE} takes the following form (Browne, 1982):

$$F_{\text{MLE}} = \log|\Sigma(\boldsymbol{\theta})| + tr\left(\mathbf{S} * \Sigma^{-1}(\boldsymbol{\theta})\right) - \log|\mathbf{S}| - p, \tag{4}$$

where for p observed variables, F_{MLE} quantifies the difference between the model-implied covariance matrix $\Sigma(\boldsymbol{\theta})$ and the sample covariance matrix $\mathbf{S}$, with $\boldsymbol{\theta}$ containing model parameters.

For the penalty term $\lambda P(\cdot)$ in Eq. 3, the regularization parameter λ controls the amount of penalty, ranging from 0 to infinity. If λ = 0, F_{RegSEM} is equal to F_{MLE}. A greater λ penalizes parameters more heavily toward 0, resulting in a simpler model with fewer parameters. The λ value can be assessed using bootstrapping or cross-validation, within sample, or determined by a fit measure. $P(\cdot)$ is a function for summing model parameters. This function should only include parameters that are to be penalized such as cross-loadings (e.g.,

detect factor loading patterns) and/or path coefficients (e.g., detect measurement bias using RFA or MIMIC models). Two common forms of $P(\cdot)$ are the L_1-norm: $\| \boldsymbol{\theta} \|_1 = \sum_j |\theta_j|$, and the L_2-norm: $\| \boldsymbol{\theta} \|_2 = \sqrt{\sum_j \theta_j^2}$, where θ_j denotes the j^{th} parameter.

The *ridge* regularization (Hoerl & Kennard, 1970) applies the L_2-norm penalty. The ridge shrinks parameter values toward 0 but not exactly 0, because the sum squares of parameter values lead to slow penalization. The ridge is less optimal for variable selection, but it retains the correlated features among variables and could improve model estimation.

The *least absolute shrinkage and selection operator* (*lasso*; Tibshirani, 1996) uses the L_1-norm penalty that penalizes the sum of absolute parameter values. The lasso can set noise parameters directly to zero, making it more suitable for variable selection. Meanwhile, signal parameters are also likely to be shrunk heavily toward zero.

The *elastic net* (*enet*; Zou & Hastie, 2005) encompasses characteristics from both the ridge (handles collinearity) and the lasso (performs variable selection) through controlling an additional parameter α ($\alpha \in [0, 1]$). The enet penalty is:

$$P_{\text{enet},\alpha}(\boldsymbol{\theta}) = (1 - \alpha) \| \boldsymbol{\theta} \|_2 + \alpha \| \boldsymbol{\theta} \|_1. \quad (5)$$

When α is 0 or 1, the enet becomes the ridge or lasso estimator, respectively. For $\alpha \in (0, 1)$, the enet is a convex combination of the lasso and ridge penalty. The enet maintains the feature of lasso for variable selection, while overcoming the over-shrinkage of non-zero parameters by regularizing more similarly to ridge. The enet is useful for variable selection (lasso) from a set of correlated variables (ridge).

The *adaptive lasso* (*alasso*; Zou, 2006) assigns different weights to re-scale parameters that are measured on different score scales. One approach for weighting is using the unpenalized MLE estimates $\boldsymbol{\theta}_{\text{MLE}}$ as the weights, leading to the following penalty function:

$$P_{\text{alasso}}(\boldsymbol{\theta}) = \| \boldsymbol{\theta}_{\text{MLE}}^{-1} * \boldsymbol{\theta} \|_1. \quad (6)$$

The alasso may overcome the limitations of the lasso by using weights. The lasso gives equal penalization to parameters on different scales, potentially producing inconsistent parameter estimates.

Practical considerations

In psychological research, survey data may contain a large number of variables in a small sample due to concerns of the time and costs. In a review of 194 studies including 1,409 confirmatory factor analysis models by Jackson, Gillaspy, and Purc-Stephenson (2009), 25% of the models consisted of more than 24 observed indicators and 20.3% of the models used sample sizes less than 200. When the number of observations per variable is small, only limited information can be provided during the model estimation, which may incur computational issues in traditional SEM, leading to model non-convergence, and biased estimates in chi-square statistics, fit measures, model parameters and standard errors (Moshagen, 2012; Shi, Lee, & Maydeu-Olivares, 2019).

Whilst SEM is generally viewed as a confirmatory methodology with the model structure specified *a priori*, regularization methods are conceptualized as incorporating exploratory features into the confirmatory models. As part of the model becomes exploratory, regularization is often used to reduce model complexity and decrease the variability of parameter estimates. When the sample size is small relative to the number of variables, the variances of parameter estimates tend to be high. At the time regularization attempts to reduce the variance, it will also bias the coefficient estimates toward zero. Regularization explores the bias-variance trade-off (Friedman, 1997). In practice, it could be worth accepting a little bias and model misfit, in return for a decrease in variances (in reference to the estimator bias-variance trade-off) and for a parsimonious model. This property of regularization is desirable when the model is sparse with few important parameters/variables to be selected (see Jacobucci et al., 2019).

Review of measurement bias detection via regularization

The early extension of regularization to SEM for measurement invariance assessments was within the Bayesian SEM (BSEM) framework (Muthén & Asparouhov, 2012, 2013). The idea is to assign small-variance normal distribution priors to parameter differences between paired groups. Thus, parameter differences are regularized toward zero unless the data likelihood indicates strong evidence against the shrinkage. Results show that BSEM is capable of detecting non-invariant parameters, but only works well when parameter differences follow the prior distribution. BSEM as a variable-selection method can also guide model re-specifications to arrive at partial invariance models. Shi, Song, Liao, Terry, and Snyder (2017) proposed using BSEM to first select a reference indicator and second detect non-invariant parameters. They found that BSEM produced lower power and lower Type I errors than the MLE likelihood ratio test with sample sizes less than 200. Moreover, BSEM has been applied to evaluate the impact of longitudinal approximate invariance on estimating structural coefficients in autoregressive cross-lagged models (Liang, Yang, & Huang, 2018). The structural coefficient estimates were affected by varying design factors and the prior choices. Findings called for further research on more efficient priors such as the G-L priors.

Ridge regularization is a likelihood-based counterpart and equivalent to BSEM with normal priors in both contexts of regression (Park & Casella, 2008) and SEM (Jacobucci & Grimm, 2018; Lu, Chow, & Loken, 2016). As the reviews discussed above, both the ridge and Bayesian ridge are less appropriate for measurement bias detection. Huang (2018) proposed using penalized likelihoods with multi-group SEM for measurement invariance evaluation. Similar to the idea of BSEM, parameter differences between groups are treated as penalized parameters and given different penalties. Specifically, the author investigated the minimax concave penalty (MCP; Zhang, 2010), of which lasso is a special case, and found that MCP was efficient in detecting measurement bias. This conclusion was confirmed by Finch (2018), who compared MLE, lasso and MCP, and concluded that MCP yielded a good balance between true

and positive rates. Lindstrøm and Dahl (2019) examined the lasso for penalizing the parameter difference in two-group SEM, and cautioned that the penalized procedure, though providing exploratory results, should not replace the traditional hypothesis testing. These findings were also consistent with Tutz and Schauberger (2015) and Magis, Tuerlinckx, and De Boeck (2015), who proposed the use of penalized likelihoods in item response theory (IRT) models to identify differential item functioning (DIF) items, or uniform biased items.

None of the studies that use regularization methods for measurement bias detection have used the RFA or MIMIC models with a continuous violator. Investigations have not been conducted on comparing the alasso and enet to the lasso and MLE. In the current study, the performance of regularization methods in comparison to MLE for measurement bias detection was explored through a simulation study and an empirical example.

Study 1: Simulation

Design factors for data generation

Factor structure

The data generation model contained one latent factor η measured by 16, 48, or 80 continuous items. The number of items chosen was based on a review of empirical research (Hoogland & Boomsma, 1998) and also consistent with simulation studies investigating the impact of model sizes (Moshagen, 2012; Shi et al., 2019). The model also included an exogenous-observed covariate ν introducing uniform bias, and an interaction between η and ν ($\eta^*\nu$) introducing non-uniform bias in items. A non-zero path from ν or $\eta^*\nu$ to an item indicates the presence of uniform or non-uniform bias, respectively. The η and ν were generated following a normal distribution $N(0,1)$. Interaction scores $\eta^*\nu$ were generated by taking the product of η and ν. The correlation between ν and η was fixed at 0. All factor loadings λ were set to .7. Residual variances were set to $(1-\lambda^2)$, so non-biased items were standardized to follow $N(0,1)$.

Proportion of biased items

The proportion of biased items was 25% or 50%, representing medium or high contamination in empirical settings. These proportions are common in practice and consistent with previous simulation studies in measurement invariance (e.g., French & Finch, 2008; Shi et al., 2017).

Condition of bias

Three conditions of bias were considered: (1) uniform bias only, (2) non-uniform bias only, and (3) both uniform and non-uniform biases. The first 25% or 50% of the items were generated as the biased items. The sizes for uniform bias (path from ν to x) or non-uniform bias (path from $\eta^*\nu$ to x) were set to: 0, .176, and .333, accounting for 0%, 3%, and 10% of the total variance in an item. The three sizes of bias were referred to as no bias, small bias, and large bias, respectively (Jak, Oort, & Dolan, 2013). The combination of uniform and non-uniform biases resulted in the total bias in an item ranging from 0% to 18.15%.

In uniform bias only conditions, the first half of the biased items had medium bias (.176), and the second half had large bias (.333). In non-uniform bias only conditions, the first, second, third, and fourth quarters of the biased items had bias of .176, .333, .176, and .333, respectively. In conditions with both biases, there were four combinations of uniform and non-uniform biases: (.176, .176), (.176, .333), (.333, .176), and (.333, .333), for the first, second, third, and fourth quarters of the biased items, respectively. For all invariant items (no bias with respect to ν), the path coefficients from ν and/or $\eta^*\nu$ to these invariant items were given 0 during data generation.

Sample size

Four sample sizes were examined: 100, 200, 500 and 1000, resulting in the ratio of the sample size (n) over the number of items (p) ranging from 1.25 to 62.5. Studies on traditional SEM have raised different recommendations on the ratio of n:p; for example, Bentler and Chou (1987) suggested a ratio of 5:1 for normally distributed data and 10:1 for data with other distributions. Although previous ad-hoc recommendations varied depending on the simulation designs, a n:p ratio as low as 1.25 in our study is considered very small and likely to cause estimation issues using traditional SEM (Deng et al., 2018). The wide range of n:p ratios used in the current study allows for investigating the benefits of regularization over MLE in SEM, and reflects typical sample sizes in empirical research.

Accordingly, this simulation study included 72 data generation conditions: 3 numbers of items (16, 48, 80) × 2 proportions of biased items (.25, .50) × 3 conditions of bias (uniform bias only, non-uniform bias only, both biases) × 4 sample sizes (100, 200, 500, 1000).

Data analysis

Two hundred datasets were generated for each condition using R (R Core Team, 2019). Data were analyzed using the lavaan package (Rosseel, 2012) for MLE, and the regsem package (Jacobucci et al., 2016) for the lasso, alasso, and enet. The analysis model is shown in Figure 1. The latent factor η was identified by fixing its variance to 1. The interaction of $\eta^*\nu$ was computed during data generation and considered as a covariate during data analysis.[1] All indicators x_1 to x_p are regressed on ν and/or $\eta^*\nu$. It is anticipated that path coefficients (ν and/or $\eta^*\nu$ to x) were estimated as non-zero when items were biased, and regularized to zero when items were unbiased. For each dataset, 15 models were tested with 15 λ (regularization parameter) values[2], starting from 0 with a .05 increment. In addition, for the enet, parameter α was fixed at .5, weighting the ridge and lasso equally. The

[1] In empirical data analysis, because latent factor scores are unknown, it is not possible to compute the interaction scores between the violator and the latent factor. We discuss how to empirically estimate the latent interaction effect in more detail in the Discussion.

[2] In practice, it is recommended to test more λ values (e.g., 40). However, the computational speed limited our ability to test more λ values in the current study.

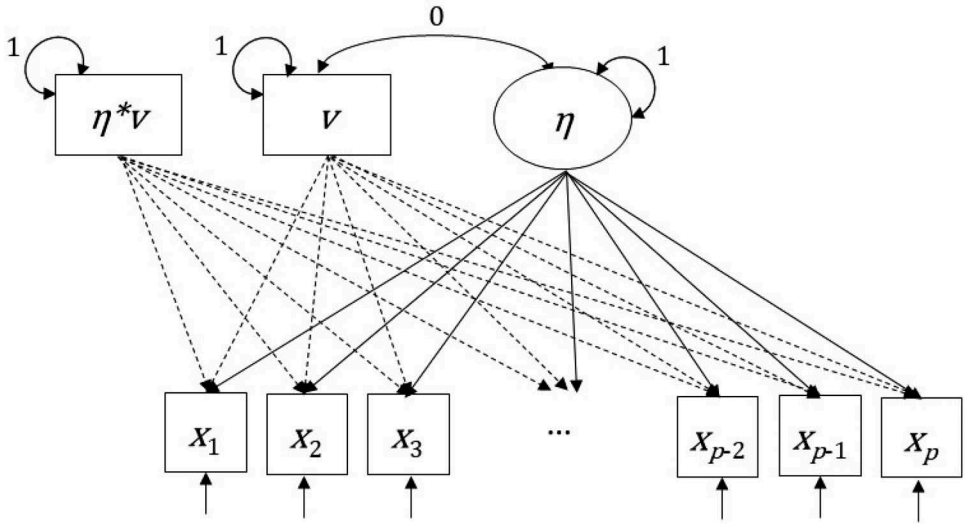

Figure 1. Restricted factor analysis model for data analysis.

best-fitting model was selected according to the smallest Bayesian information criterion (BIC; Schwarz, 1978) value.

Outcomes were evaluated based on true positive rates (proportion of replications that correctly detect measurement bias for biased items) and false positive rates (proportion of replications that falsely conclude bias for unbiased items). For MLE, a Wald test at the .05 alpha level was used to determine the significance of the bias estimates (v and/or η^*v to x). Items with significant non-zero path coefficients were labeled as having measurement bias. For the lasso, alasso, and enet, standard errors of parameter estimates are difficult to obtain and not available in regsem, and thus hypothesis testing was not conducted. Items were simply labeled as biased if path coefficient estimates were non-zero, even with nearly zero values. Further, parameter estimates were evaluated using the root mean square errors (RMSE): $\mathrm{RMSE}\left(\hat{\theta}\right) = \sqrt{\sum_{r=1}^{n_r}\left(\hat{\theta}_r - \theta\right)^2 / n_r}$, where $\hat{\theta}_r$ and θ are the parameter estimate at the r^{th} replication and the population parameter value, respectively, and n_r is the number of replications. RMSE reflects both the estimation bias (deviation from the population value) and sampling variability (variability of parameter estimates).

Results

Results from each condition were aggregated over the replications and across the items with the same type and size of bias. We only present results for uniform biased items and non-biased items in bias conditions 1 (items with uniform bias only) and 3 (items with both biases). Results for non-uniform biased items were similar to those for uniform biased items.

True positive rates

Figures 2 and 3 show the true positive rates for detecting uniform biased items in bias conditions 1 and 3, respectively. Overall, true positive rates increased as the size of bias increased, sample size increased, and percent of biased items reduced. The impact of the number of items on the true positive rates was not notable. When items only presented uniform bias (Figure 2), the enet yielded the highest true positive rates except in conditions where 50% items had small bias of .176. In these conditions, MLE produced the greatest true positive rates. For conditions with 50% biased items in small samples (<200), the lasso yielded the lowest true positive rates. When items presented both uniform and non-uniform biases (Figure 3), the enet produced the highest true positive rates in nearly all conditions. The lowest rates were mostly associated with MLE in sample sizes of less than 200. The lasso often yielded higher true positive rates than the alasso and MLE in conditions with sample sizes greater than 100 or 200.

False positive rates

Figure 4 shows the false positive rates for incorrectly concluding bias in non-biased items. The alasso well controlled the false-positive rates in the uniform bias only condition (condition 1). However, when 50% items had both uniform and non-uniform biases (condition 3), false positive rates were substantially inflated; with sample sizes greater than 200, the false positive rates of the alasso increased rapidly as sample sizes increased (up to .98). The enet yielded the greatest false positive rates in nearly all conditions. Especially when the percent of biased items was 50%, the false positive rates for the enet could be elevated to near .55 in bias condition 1 and about .92 in bias condition 3, with the sample size of 1000. The lasso showed the second greatest inflation in false-positive rates. MLE produced false positive rates (around .1) slightly higher than the preset alpha level of .05 in all conditions.

It is worth repeating that MLE determines biased items using the hypothesis testing so small non-zero estimates may not indicate measurement bias unless they lie outside the 95% confidence interval. In contrast, the lasso, alasso, and enet conclude measurement bias only depending on whether path coefficient estimates (v and/or η^*v to x) were zero (no bias) or non-zero (measurement bias). This may explain why the false positive rates were inflated substantially when both biases were present in 50% of the items. With many items having both biases, the penalized parameters, though estimated near zero, may not be exactly zero, which lead to high false positive rates.

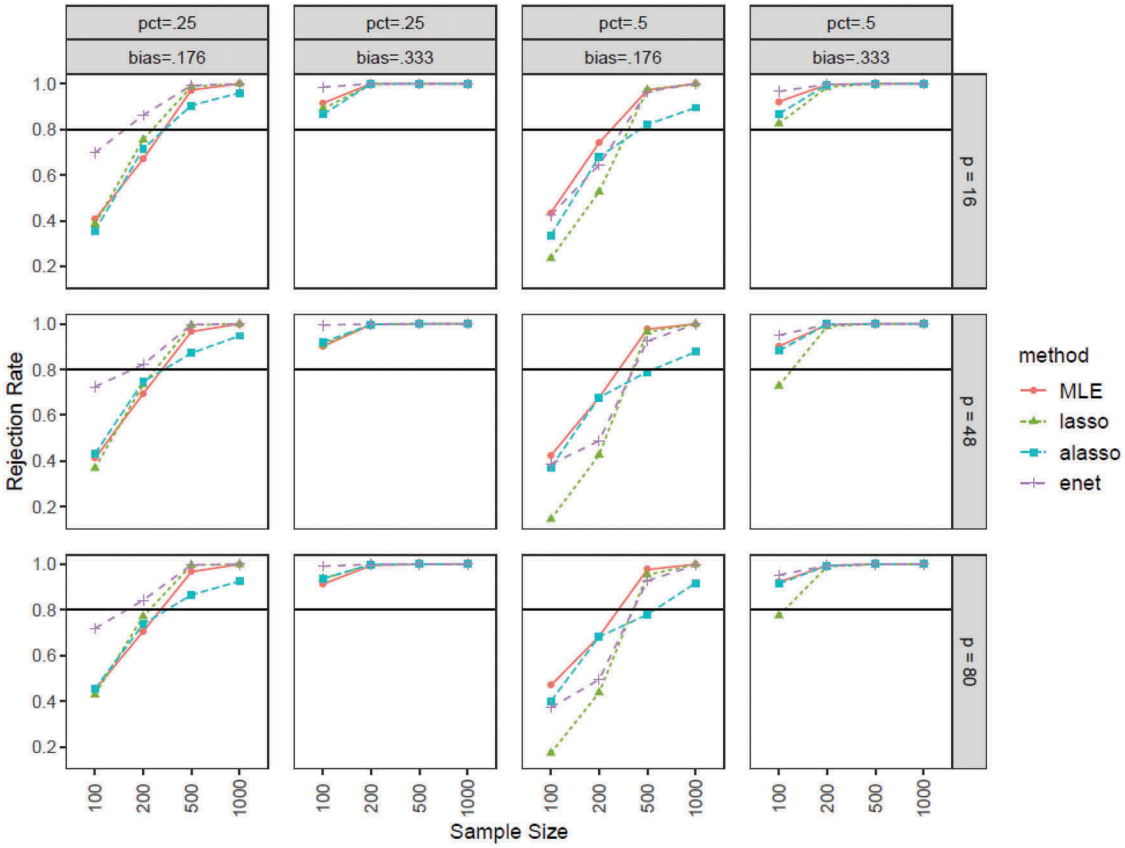

Figure 2. True positive rates for detecting uniform bias in conditions with uniform bias only.
Note. pct = percent of biased items, p = number of observed indicators.

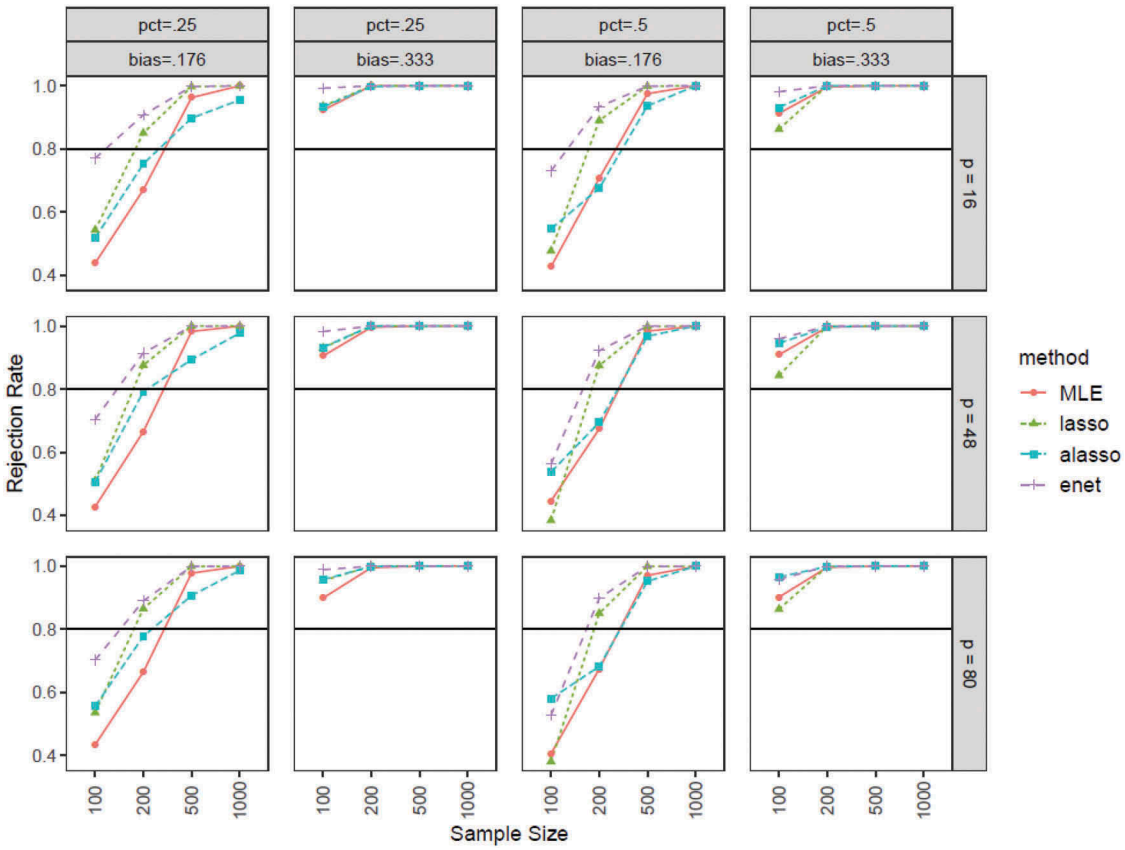

Figure 3. True positive rates for detecting uniform bias in conditions with both biases.
Note. pct = percent of biased items, p = number of observed indicators.

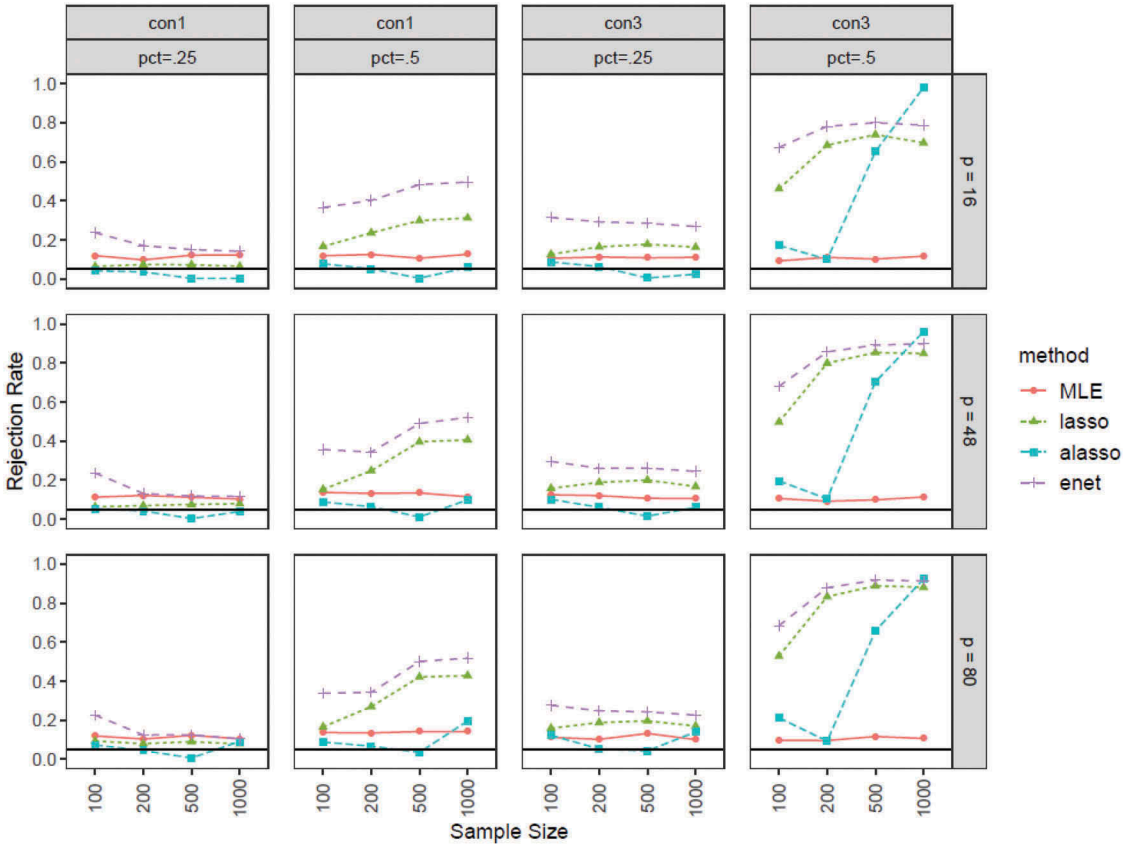

Figure 4. False positive rates for non-biased items in bias conditions 1 (uniform bias only) and 3 (both biases).
Note. pct = percent of biased items, p = number of observed indicators, con1 = bias condition 1 – uniform biased only, con3 = bias condition 3 – both biases.

RMSE

Figures 5 and 6 show the RMSE of path coefficient estimates for uniform biased items (v to x) in bias conditions 1 (uniform bias only) and 3 (both biases), respectively. RMSE appeared to be affected the most by the sample size, but not markedly by other factors. MLE yielded the lowest RMSE. When the size of bias was .176, the alasso tended to produce the greatest RMSE in most conditions. With the size of bias at .333, the largest RMSE was associated with the lasso (in smaller samples) and the enet (in larger samples).

Figure 7 shows the RMSE of path coefficient estimates for non-biased items in bias conditions 1 and 3. In general, RMSE decreased as the sample size increased. MLE yielded the greatest RMSE and was affected the most by the sample size. All regularization methods were least affected by the design factors and provided low RMSE. As regularization penalizes estimates toward or directly to zero, it was reasonable to observe relatively high RMSE for coefficient estimates associated with the biased items and low RMSE associated with the non-biased items.

Study 2: An empirical illustration

Method

In Study 2, regularization with RFA for detecting uniform bias in regard to a continuous violator was investigated using the data from Holzinger and Swineford (1939). The dataset consists of 26 psychological tests administered to 7th and 8th grade students in two schools. Nineteen out of the 26 tests were intended to measure four correlated latent factors: (1) spatial (η_1) measured by visual perception, cubes, paper form board, and flags (x_1–x_4), (2) verbal (η_2) measured by general information, paragraph comprehension, sentence completion, word classification, and word meaning (x_5–x_9), (3) speed (η_3) measured by addition, code, counting groups of dots, and straight and curved capitals (x_{10}–x_{13}), and (4) memory (η_4) measured by word recognition, number recognition, figure recognition, object-number, number-figure, and figure-word (x_{14}–x_{19}). In this simple-structure model, each test loaded on only one target factor. The age of the student was recoded as a composite of age in years (agey) and months into the current year (agem): age = agey+agem/12, and used as the violator. All test scores were standardized before the analysis.

Research has indicated that cross-loadings are possible in the four-factor model described above (Muthén & Asparouhov, 2012). As cross-loadings are unknown, the analysis model was developed to estimate all potential cross-loadings simultaneously with the examination of measurement bias. Nonetheless, allowing all potential cross-loadings freely estimated resulted in model non-convergence using MLE in the current dataset. The use of regularization on cross-loadings may reduce the number of parameters by setting their estimates to zero, and therefore could resolve the issues of model non-convergence.

Accordingly, a four-factor model was fitted to the data, with all observed indicators loaded on each factor. All latent factor

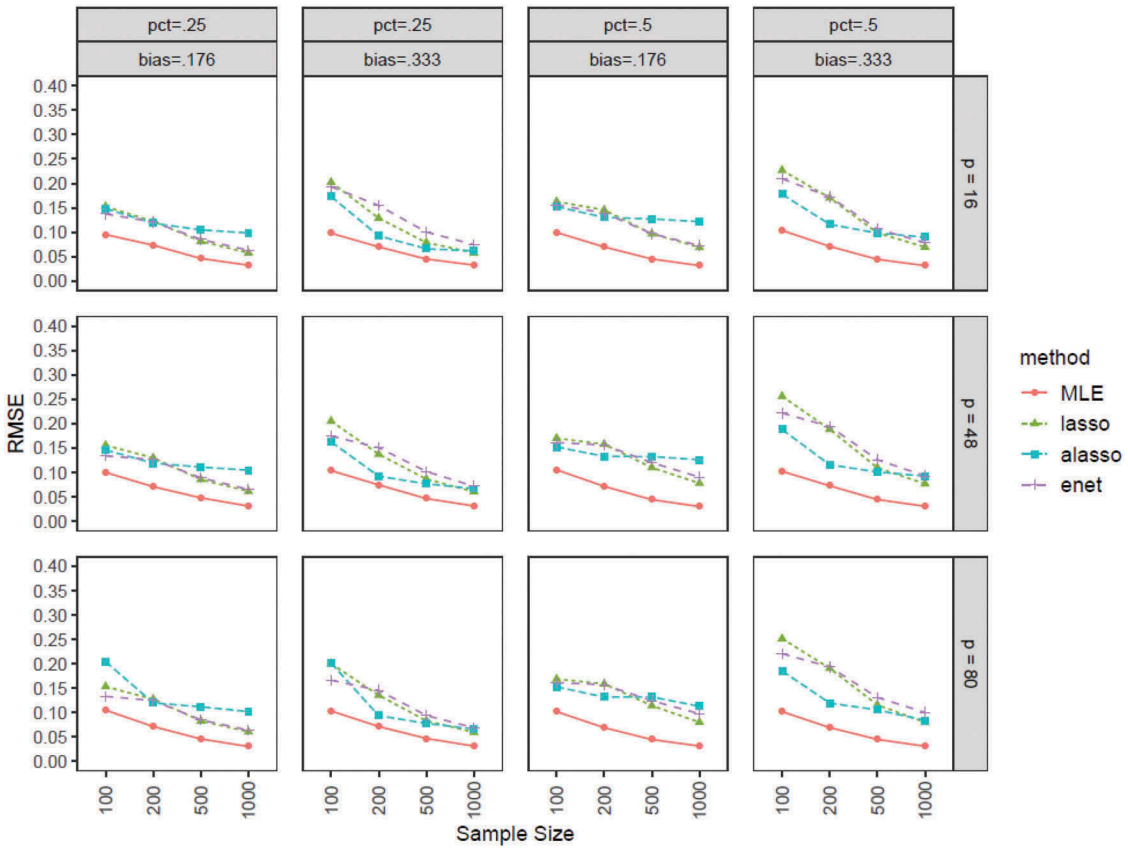

Figure 5. RMSE for estimating uniform bias in conditions with uniform bias only.
Note. pct = percent of biased items, p = number of observed indicators.

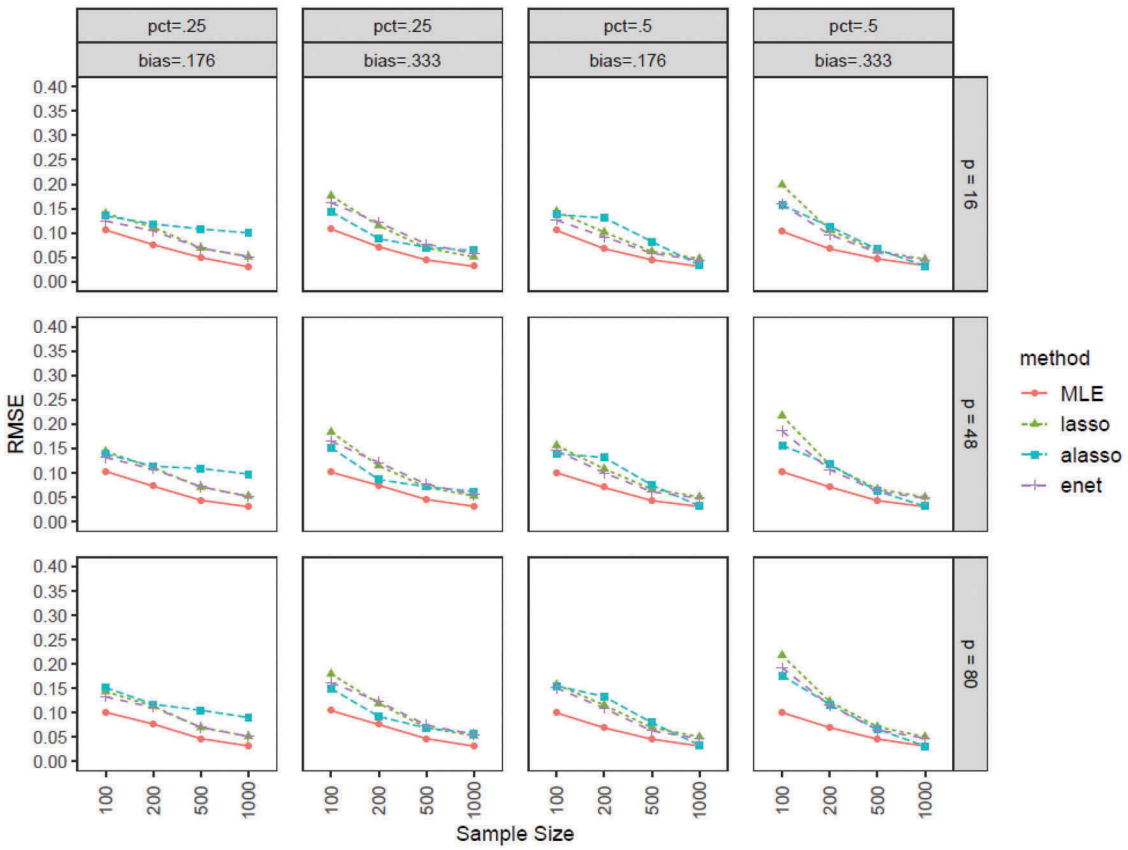

Figure 6. RMSE for estimating uniform bias in conditions with both biases.
Note. pct = percent of biased items, p = number of observed indicators.

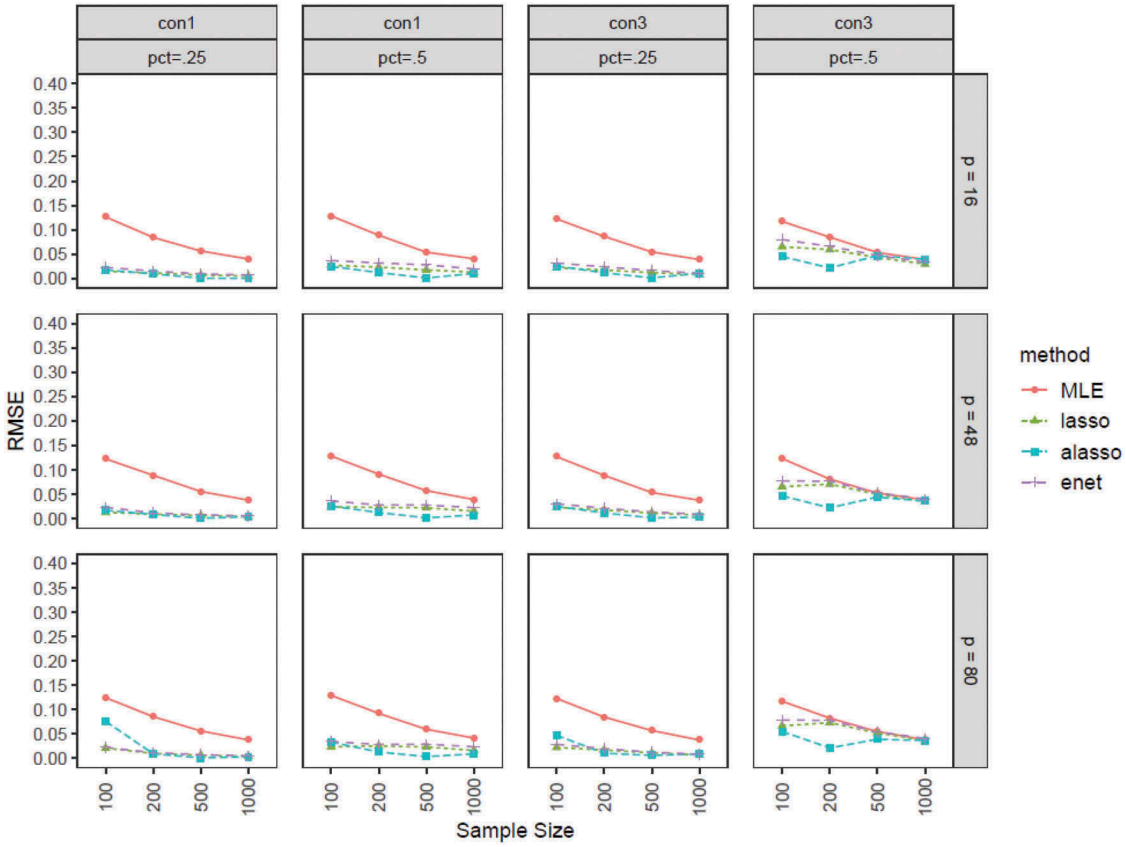

Figure 7. RMSE for estimating non-biased items in bias conditions 1 (uniform bias only) and 3 (both biases).
Note. pct = percent of biased items, p = number of observed indicators, con1 = bias condition 1 – uniform biased only, con3 = bias condition 3 – both biases.

variances were fixed at 1 for model identification. The age was regressed on all 19 indicators, and non-zero path coefficients (from age to x) indicated uniform bias. The ridge, lasso, alasso, and enet regularizations were applied to all cross-loadings and path coefficients indicating uniform bias. Although the ridge was not included in the simulation, we showed its performance in comparison with the lasso, alasso, and enet in the real data analysis. Similar to the simulations, for each regularization method, 15 models were fitted by applying 15 regularization parameter λ (starting 0 with .05 increment); moreover for the enet, the mixture parameter α was set at .5. Different from the simulations, both BIC and Akaike information criterion (AIC; Akaike, 1974) were used to guide the selection of the best fitting-regularized model. The model with the lowest BIC or AIC was selected. Root mean square error of approximation (RMSEA; Steiger & Lind, 1980) was also reported as a model fit measure.

Results

Table 1 presents the estimates for uniform bias in regard to the age using the ridge, lasso, alasso, and enet. The BIC solutions led to a more parsimonious model (more parameter estimates being penalized to zero), while AIC selected a more complex model (less penalty imposed on parameter estimates). As seen from the λ values, BIC always chose the model with a penalty (λ) equal to or greater than AIC. This is consistent with the measurement invariance literature using information criteria as model selection methods (Huang, 2018; Liang & Luo, 2019), in which BIC was shown to favor a simpler model and AIC preferred a more complex model.

For the ridge, both BIC and AIC selected the same best-fitting model (λ = .05, RMSEA = .040). All path coefficient estimates were non-zero that indicated bias in all tests with respect to age. These biases ranged from −.22 to .24. For the lasso, the model selected by BIC (λ = .50, RMSEA = .065) indicated that only one test: counting groups of dots (x_{12}), was biased across ages. The model selected by AIC (λ = .05, RMSEA = .033) suggested that 17 out of 19 tests had uniform bias, ranging from −.17 to .22; only cubes (x_2) and code (x_{11}) exhibited no bias. For both the alasso and enet, BIC solutions (for alasso: λ = .25, RMSEA = .065; for enet: λ = .70, RMSEA = .066) indicated that no tests were biased with respect to age. However, AIC solutions suggested that all tests were biased using the alasso (λ = 0, RMSEA = .040), and 18 out of 19 tests were biased using the enet (λ = .05, RMSEA = .032). It is worth noting that when λ = 0, coefficient estimates are equal to MLE estimates. The biases for alasso ranged from −.24 to .25, and for enet ranged from −.19 to .23. It appeared that AIC and RMSEA provided the same suggestion for model selection, and chose models with less penalty than BIC. All regularization methods consistently suggested that tests that exhibited the highest bias were paragraph comprehension (x_6), sentence completion (x_7), word classification (x_8), word meaning (x_9), and counting (x_{12}).

In addition, when examining the factor loading patterns based on the smallest BIC values, the ridge resulted in all except for one cross-loading (η_1 by x_6) estimated with non-zero values. The lasso and enet penalized all cross-loadings to

Table 1. Estimates for uniform bias in regard to age using Ridge, Lasso, Alasso, and Enet.

	Ridge		Lasso		Alasso		Enet	
	BIC λ = .05	AIC λ = .05	BIC λ = .50	AIC λ = .05	BIC λ = .25	AIC λ = .00	BIC λ = .70	AIC λ = .05
x_1~ Age	−.04	−.04	–	−.03	–	−.05	–	−.04
x_2~ Age	.00	.00	–	–	–	.00	–	–
x_3~ Age	.08	.08	–	.07	–	.07	–	.07
x_4~ Age	.04	.04	–	.02	–	.04	–	.03
x_5~ Age	−.13	−.13	–	−.07	–	−.14	–	−.10
x_6~ Age	−.19	−.19	–	−.14	–	−.20	–	−.17
x_7~ Age	−.22	−.22	–	−.16	–	−.24	–	−.19
x_8~ Age	−.21	−.21	–	−.17	–	−.23	–	−.19
x_9~ Age	−.16	−.16	–	−.10	–	−.17	–	−.13
x_{10}~ Age	.13	.13	–	.10	–	.13	–	.11
x_{11}~ Age	.01	.01	–	–	–	.01	–	.00
x_{12}~ Age	.24	.24	.05	.22	–	.25	–	.23
x_{13}~ Age	.11	.11	–	.09	–	.11	–	.10
x_{14}~ Age	−.03	−.03	–	−.02	–	−.03	–	−.02
x_{15}~ Age	.08	.08	–	.06	–	.08	–	.07
x_{16}~ Age	−.09	−.09	–	−.07	–	−.09	–	−.08
x_{17}~ Age	.06	.06	–	.04	–	.06	–	.05
x_{18}~ Age	.04	.04	-	.02	–	.04	–	.03
x_{19}~ Age	−.03	−.03	-	−.01	–	−.04	–	−.02

zero and led to a simple-structure four-factor model. The alasso identified one cross-loading (η_1 by x_7) being non-zero (i.e., .05), and all other cross-loadings were penalized to zero. In line with previous studies, the ridge was more likely to detect cross-loadings, whereas the lasso, alasso, and enet led to more parsimonious cross-loading structures based on the BIC solutions.

Discussion

Correct detection of measurement bias could help researchers revise models or refine psychological scales. This study compared the performance of MLE, lasso, alasso, and enet for detecting various forms of measurement biases (uniform, non-uniform and both biases) in regard to a continuous violator using RFA models. We focused on examining complex SEM models with a large number of variables relative to small sample sizes. Through a simulation study and an illustrative example, we showed that regularization methods could provide useful exploratory results that complement findings using MLE in traditional SEM modeling.

As a summary of the findings, among the three regularization methods, the enet yielded greater true positive rates than the lasso and alasso, but also resulted in more inflated false-positive rates. The alasso produced slightly greater true positive rates and lower false positive rates than the lasso when only one type of biases (uniform or non-uniform) was present. With both biases, the alasso yielded slightly lower true positive rates than the lasso, but still had a better control of false positive rates in most conditions. Compared to regularization methods, MLE performed comparably on true positive rates in most conditions, but was inferior when both biases were present in 25% of the items with small sample sizes (<200). False positive rates for MLE were slightly inflated but stable across conditions. MLE produced lower RMSE for biased items but higher RMSE for non-biased items than regularization methods. The numbers of indicators did not appear to affect the outcomes evaluated. Across all of these comparisons, there was a tradeoff between MLE and the regularization methods, which is in line with prior research (e.g., Jacobucci et al., 2019).

Accordingly, we summarized the implications and recommendations for researchers interested in applying regularization in measurement bias detection. First, if the goal is to identify biased items, the enet often performed better at smaller sample sizes (<500) and may be prioritized over other methods. Second, if one worries about false positives, the alasso yielded low false positive rates for non-biased items outside of a high number of biased items. Although in few cases noise parameters may not be penalized directly to zero by the alasso, causing relatively high false positive rates, the estimates of these parameters were still close to zero (indicated by the low RMSE in non-biased items). Third, if the goal is the overall model estimation for biased items, MLE is recommended because it produces unpenalized parameter estimates which are less biased than penalized parameter estimates.

Nevertheless, the above findings need to be interpreted with caution. First, standard errors of penalized estimates were difficult to obtain, and have not been evaluated. Bias detection based on penalized estimates does not rely on inferential procedures comparing to MLE. Therefore, regularization methods are not appropriate for hypothesis testing and may better serve in an exploratory manner for variable selection. Second, although the goal of regularization is to perform both variable selection and model estimation, for the ridge, lasso, and enet, global penalization is imposed on all parameters without controlling for the recovery of local signals. That is, both the noise and signal parameters were penalized the same amount toward zero, leading to the penalized parameter estimates for biased items deviated more from the population values than MLE unpenalized estimates. This may be one reason for observing greater RMSE in penalized estimates for the biased items (though the variability of penalized parameters may reduce). For non-biased items, given that many parameters were penalized to nearly zero or zero, RMSE was reasonably lower using regularization methods than MLE. Third, it should note that the use of BIC as a model selection criterion tended to select a more parsimonious

model, and thus potentially lower the power of detecting bias. As shown in our empirical example, AIC marked more items as biased than BIC. Also confirmed from Huang (2018), BIC yielded lower true positive rates than AIC. Using a different model selection criterion may yield a different result pattern.

Given the inconsistent recommendations from different information criteria, one consideration would be to use a proper effect size measure of bias in addition to the evaluation of true and false positives. One measure of the effect size for measurement equivalence was developed at the item level (Nye, Bradburn, Olenick, Bialko, & Drasgow, 2019; Nye & Drasgow, 2011). These studies show that the proposed effect size can complement significance testing; however, this effect size can only apply to evaluate bias between two groups. Another effect size measure that can accommodate continuous violators was developed under MIMIC models (Jin, Myers, Ahn, & Penfield, 2013), but it can only be used for uniform bias. Developing effect size measures for both uniform and non-uniform biases with respect to continuous violators would be a useful direction for future research. Especially in regularized SEM where parameter estimates are penalized to varying degrees, developing a meaningful effect size would facilitate the understanding of the practical importance of measurement bias.

The real data analysis in this study gives an example in which MLE did not provide a convergent solution while regularization was able to handle complex models by forcing a number of parameter estimates to zero. Model non-convergence was not observed in the simulation study probably because it used a rather simple one-factor model. The empirical study examined a complex four-factor model with 19 indicators, estimating all cross-loadings along with the estimation for bias. In practice, when a model is not estimable using MLE in traditional SEM, regularization can be considered as a viable alternative to explore the important parameters, eliminate noise parameters, and achieve a simpler model structure.

Although regularization reduces parameter variances at the expense of increasing estimations bias, the more important goal in measurement invariance research is to achieve generalization by filtering out trivial measurement bias while focusing on the detection of non-trivial bias. For example, during the scale development, researchers may be more interested in removing or revising a few items with non-ignorable bias, instead of reviewing many items with trivial bias. If for practical reasons some biased items need to be retained in the test, understanding items with non-trivial bias will inform incoming research using the same test, and benefit future modeling on data containing non-trivial biased items.

Limitations and future research

This study has a number of limitations which can be considered as future research directions. In the empirical study using the Holzinger and Swineford data, the interaction between the latent factor and the observed violator for examining non-uniform bias was not modeled. Although latent interactions are relatively easy to model in M*plus* (Muthén & Muthén, 1998–2010) using the XWITH option with latent moderated structural equations, modeling the latent interactions is less straightforward in lavaan. With lavaan, one way to model the latent interactions is to use the product indicators approach. For example, the observed violator can be modeled as a latent factor with one indictor (e.g., v_1). A set of product indicators (e.g., $x_1{}^*v_1$, $x_2{}^*v_1$, $x_3{}^*v_1$, $x_4{}^*v_1$) are then created and used to measure the *interaction* latent factor. Comparing methods for modeling latent interactions to detect non-uniform bias would be an interesting extension to the current study.

Despite the numerous benefits of regularization, these methods may not be suitable for models in which the number of indicators is greater than the sample size (Jacobucci et al., 2019). Also, considering the difficulty of obtaining standard errors of parameter estimates, an alternative approach would be to consider implementing regularization in the Bayesian framework. Recent development has extended various Bayesian regularization in SEM, such as the use of the spike and slab prior (SSP; Mitchell & Beauchamp, 1988) for determining cross-loadings patterns (Lu et al., 2016) and conducting SEM model selection (Lu, Chow, & Loken, 2017), the adaptive Bayesian lasso approach to identify multiple linear and non-linear effects in SEM (Brandt, Cambria, & Kelava, 2018), and the Bayesian adaptive lasso in multivariate generalized latent variable models (Feng, Wu, & Song, 2017a, 2017b). A didactic paper by Jacobucci and Grimm (2018) also provides an overview of the likelihood-based and Bayesian regularization extended to SEM.

Furthermore, another limitation of regularized SEM is that multiple models with varying levels of penalty need to be tested. With Bayesian estimation, hierarchical Bayesian approaches can be used which assign hyper-prior distributions on the hyperparameters of the prior distributions. The use of hyper priors allows for the test of a range of penalty in one step and has the theoretical advantage for simultaneous estimation and variable selection. Priors in this class include adaptive lasso and horseshoe priors (Carvalho et al., 2010), based on which other modern priors have also been developed. Investigating these modern priors in Bayesian regularization for detecting measurement bias warrants further work.

ORCID

Xinya Liang http://orcid.org/0000-0002-2453-2162

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Barrett, P. T., & Kline, P. (1981). The observation to variable ratio in factor analysis. *Personality Study & Group Behaviour*, *1*, 23–33.

Bentler, P. M., & Chou, C.-P. (1987). Practical issues in structural modeling. *Sociological Methods & Research*, *16*, 78–117.

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: A Multidisciplinary Journal*, *25*, 946–960.

Browne, M. W. (1982). Covariance structures. In D. M. Hawkins (Ed.), *Topics in applied multivariate analysis* (pp. 72–141). Cambridge, MA: Cambridge University Press.

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, *97*, 465–480.

Chin, W. W., Marcolin, B. L., & Newsted, P. R. (2003). A partial least squares latent variable modeling approach for measuring interaction effects: Results from a Monte Carlo simulation study and an electronic-mail emotion/adoption study. *Information Systems Research*, *14*, 189–217.

Deng, L., Yang, M., & Marcoulides, K. M. (2018). Structural equation modeling with many variables: A systematic review of issues and developments. *Frontiers in Psychology*, *9*, 580.

Fan, J., & Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, *32*, 928–961.

Feng, X.-N., Wu, H.-T., & Song, X.-Y. (2017a). Bayesian adaptive lasso for ordinal regression with latent variables. *Sociological Methods & Research*, *46*, 926–953.

Feng, X.-N., Wu, H.-T., & Song, X.-Y. (2017b). Bayesian regularized multivariate generalized latent variable models. *Structural Equation Modeling*, *24*, 341–358.

Finch, H. (2018). Comparison of measurement invariance testing using penalized likelihood and maximum likelihood estimators: A Monte Carlo simulation study. *General Linear Model Journal*, *44*, 20–33.

French, B. F., & Finch, W. H. (2008). Multigroup confirmatory factor analysis: Locating the invariant referent sets. *Structural Equation Modeling*, *15*, 96–113.

Friedman, J. H. (1997). On bias, variance, 0/1—loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, *1*(1), 55–77.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2 ed.). New York, NY: Springer.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*, 55–67.

Holzinger, K. J., & Swineford, F. (1939). A study in factor analysis: The stability of a bi-factor solution. In *Supplementary educational monographs* (Vol. 48). Chicago, IL: Department of Education, University of Chicago.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research*, *26*, 329–367.

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

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

Jackson, D. L., Gillaspy, J. A., Jr, & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods*, *14*, 6–23.

Jacobucci, R., Brandmaier, A. M., & Kievit, R. A. (2019). A practical guide to variable selection in structural equation modeling by using regularized multiple-indicators, multiple-causes models. *Advances in Methods and Practices in Psychological Science*, *2*, 55–76.

Jacobucci, R., & Grimm, K. J. (2018). Comparison of frequentist and Bayesian regularization in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*, 639–649.

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

Jak, S., Oort, F. J., & Dolan, C. V. (2013). A test for cluster bias: Detecting violations of measurement invariance across clusters in multilevel data. *Structural Equation Modeling*, *20*, 265–282.

Jin, Y., Myers, N. D., Ahn, S., & Penfield, R. D. (2013). A comparison of uniform DIF effect size estimators under the MIMIC and Rasch models. *Educational and Psychological Measurement*, *73*, 339–358.

Klein, A., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika*, *65*, 457–474.

Liang, X., & Luo, Y. (2019). A comprehensive comparison of model selection methods for testing factorial invariance. *Structural Equation Modeling*. doi:10.1080/10705511.2019.1649983

Liang, X., Yang, Y., & Huang, J. (2018). Evaluation of structural relationships in autoregressive cross-lagged models under longitudinal approximate invariance: A Bayesian analysis. *Structural Equation Modeling*, *25*, 558–572.

Lindstrøm, J. C., & Dahl, F. A. (2019). Model selection with lasso in multi-group structural equation models. *Structural Equation Modeling*, 1–10. doi:10.1080/10705511.2019.1638262

Lu, Z. H., Chow, S. M., & Loken, E. (2016). Bayesian factor analysis as a variable-selection problem: Alternative priors and consequences. *Multivariate Behavioral Research*, *51*, 519–539.

Lu, Z.-H., Chow, S.-M., & Loken, E. (2017). A comparison of Bayesian and frequentist model selection methods for factor analysis models. *Psychological Methods*, *22*, 361–381.

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.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*, 127–143.

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

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

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

Millsap, R. E., & Kwok, O.-M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods*, *9*, 93–115.

Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, *83*, 1023–1032.

Moshagen, M. (2012). The model size effect in SEM: Inflated goodness-of-fit statistics are due to the size of the covariance matrix. *Structural Equation Modeling*, *19*, 86–98.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*, 557–585.

Muthén, B. O., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*, 313–335.

Muthén, B. O., & Asparouhov, T. (2013). BSEM measurement invariance analysis. *Mplus web notes 17*.

Muthén, L. K., & Muthén, B. O. (1998-2010). *Mplus user's guide (Version 6th)*. Los Angeles, CA: Muthén & Muthén.

Nye, C. D., Bradburn, J., Olenick, J., Bialko, C., & Drasgow, F. (2019). How big are my effects? Examining the magnitude of effect sizes in studies of measurement equivalence. *Organizational Research Methods*, *22*, 678–709.

Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. *Journal of Applied Psychology*, *96*, 966–980.

Oort, F. J. (1992). Using restricted factor analysis to detect item bias. *Methodika*, *6*, 150–166.

Park, T., & Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, *103*, 681–686.

Polson, N. G., & Scott, J. G. (2010). Shrink globally, act locally: Sparse Bayesian regularization and prediction. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, & M. West (Eds.), *Bayesian statistics 9* (pp. 501–538)). Oxford, UK: Oxford University Press.

R Core Team. (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. *Journal of Statistical Software*, *48*, 1–36.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

Serang, S., Jacobucci, R., Brimhall, K. C., & Grimm, K. J. (2017). Exploratory mediation analysis via regularization. *Structural Equation Modeling*, *24*, 733–744.

Shi, D., Lee, T., & Maydeu-Olivares, A. (2019). Understanding the model size effect on SEM fit indices. *Educational and Psychological Measurement*, *79*, 310–334.

Shi, D., Song, H., Liao, X., Terry, R., & Snyder, L. A. (2017). Bayesian SEM for specification search problems in testing factorial invariance. *Multivariate Behavioral Research*, *52*, 430–444.

Steiger, J. H., & Lind, J. C. (1980). *Statistically based tests for the number of common factors*. Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

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

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

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, *38*, 894–942.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*, 1418–1429.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical methodology)*, *67*, 301–320.