Are Factor Scores Measurement Invariant?

Mark H. C. Lai[1] and Winnie W.-Y. Tse [1]

[1]Department of Psychology, University of Southern California

Author Note

Mark H. C. Lai https://orcid.org/0000-0002-9196-7406

Winnie W.-Y. Tse https://orcid.org/0000-0001-5175-6754

The preprint of this manuscript has been posted at https://osf.io/preprints/psyarxiv/uzrak. This research is based on work supported by the National Science Foundation under Grant No. 2141790.

Correspondence concerning this article should be addressed to Mark H. C. Lai, Department of Psychology, University of Southern California, 3620 McClintock Ave, Los Angeles, CA 90089-0068, Email: hokchiol@usc.edu

Abstract

There has been increased interest in practical methods for integrative analysis of data from multiple studies or samples, and using factor scores to represent constructs has become a popular and practical alternative to latent variable models with all individual items. Although researchers are aware that scores representing the same construct should be on a similar metric across samples—namely they should be measurement invariant—for integrative data analysis, the methodological literature is unclear whether factor scores would satisfy such a requirement. In this note, we show that even when researchers successfully calibrate the latent factors to the same metric across samples, factor scores—which are *estimates* of the latent factors but not the factors themselves—may not be measurement invariant. Specifically, we prove that factor scores computed based on the popular regression method are generally not measurement invariant. Surprisingly, such scores can be noninvariant even when the items are invariant. We also demonstrate that our conclusions generalize to similar shrinkage scores in item response models for discrete items, namely the expected a posteriori scores and the maximum a posteriori scores. Researchers should be cautious in directly using factor scores for cross-sample analyses, even when such scores are obtained from measurement models that account for noninvariance.

Keywords: factor scores, measurement invariance, multiple-group analysis, shrinkage

Are Factor Scores Measurement Invariant?

With an abundance of methodological research in the past few decades (e.g., Horn & McArdle, 1992; Luong & Flake, 2022; Meredith, 1964; Millsap, 2011; Widaman & Reise, 1997), social and behavioral scientists generally acknowledge that at least some degree of measurement invariance is needed for comparing observed scores (e.g., scale scores and sum scores) across groups. While full invariance is commonly violated, when only a subset of items is noninvariant and the others are invariant, one can fit a partial invariance model to obtain valid group comparisons on the latent parameters (e.g., Byrne et al., 1989; Lai et al., 2019; Shi et al., 2019). One can further obtain inferences on latent variables across groups by joint estimation of the structural parameters of the latent variables (e.g., group means, linear associations among latent variables) and the measurement parameters with the necessary full or partial invariance constraints.

The joint estimation approach, however, typically requires a model with many parameters (e.g., when there are many groups or many items), and as a result, some researchers have proposed and adopted an alternative strategy by first computing estimated factor scores as proxies for the latent variables, and then fitting a structural model among those factor score variables. For example, Curran and Hussong (2009) proposed the integrative data analysis framework for harmonizing measures from different studies and samples, and they recommended computing factor scores from each sample for subsequent inferences. Factor scores are also useful for making decisions at the individual level, such as for selecting top-ranked individuals on aptitude tests or contributing to the diagnosis of mental conditions (Lai et al., 2019; Millsap & Kwok, 2004).

However, there seems to be a conception, among both applied researchers and methodologists, that one can improve upon sum scores by computing factor scores from a partial invariance model, as the factor scores will be free of systematic measurement bias and can be validly compared across groups. For example, when Curran and Hussong (2009) recommended researchers assess measurement invariance across samples and then

calculate person scores "by using one of several available factor score estimates in the factor model" (p. 97; see also Bauer & Curran, 2016; Curran et al., 2014, 2016; Davoudzadeh et al., 2020); subsequent applied research generally follows such a practice of using factor scores after invariance testing (e.g., Luningham et al., 2019; MacDonald & Park, 2022; Zhao et al., 2022). McNeish (2022), on discussing a potential benefit of using factor scores over sum scores, suggested that "it would not make sense to compare sum scores across populations" when invariance is violated, but "[F]actor scoring can address some of these issues" by "allowing partial measurement invariance." (p. 4281). Steenkamp and Maydeu-Olivares (2021) similarly suggested that sum scores require measurement invariance, but "factor scores based on the (partial) scalar invariance model" (p. 14) would allow for cross-country comparisons.

However, as demonstrated in this short note, factor scores obtained by the popular *regression* method (Thomson, 1935), which was used in both McNeish (2022) and Steenkamp and Maydeu-Olivares (2021), are not necessarily invariant, even when they are obtained from a partial invariance model where the noninvariance is adjusted so that the latent variables can be compared. Perhaps even more surprisingly, such factor scores, hereafter *regression factor scores*, may *not* be invariant even when invariance holds for the items. We will use two heuristic examples with simulated data to illustrate the points. The mathematical details then follow. In addition, we also illustrate that such properties of the regression factor scores also generalize to similar scores based on item response models, namely the expected a posteriori (EAP) and the maximum a posteriori (MAP) scores. The code behind all analyses has been made publicly available on GitHub and can be accessed at https://github.com/marklhc/fsinv-supp/. This study was not preregistered.

Example 1: When Invariance Holds

In the first example, we simulate data with two groups, each of size 50, and three items based on the common factor model

$$\mathbf{y}_g = \boldsymbol{\nu}_g + \boldsymbol{\lambda}_g \eta + \boldsymbol{\varepsilon}_g, \tag{1}$$

where $g = 1, 2$ indexes groups, η is the latent factor, and invariance holds such that the two groups have equal factor loadings ($\boldsymbol{\lambda}_1 = \boldsymbol{\lambda}_2$), intercepts ($\boldsymbol{\nu}_1 = \boldsymbol{\nu}_2$), and unique variances and covariances ($V[\boldsymbol{\varepsilon}_1] = \boldsymbol{\Theta}_1 = \boldsymbol{\Theta}_2 = V[\varepsilon_2]$). However, the factor mean ($\alpha = E[\eta]$) is higher for Group 2, whereas the factor variance ($\psi = V[\eta]$) is higher for Group 1. We use the following parameter values:

$$\boldsymbol{\lambda}_1 = \boldsymbol{\lambda}_2 = \begin{bmatrix} 0.9 \\ 0.7 \\ 0.5 \end{bmatrix}, \boldsymbol{\nu}_1 = \boldsymbol{\nu}_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \boldsymbol{\Theta}_1 = \boldsymbol{\Theta}_2 = \begin{bmatrix} 0.19 & & \\ 0 & 0.51 & \\ 0 & 0 & 0.75 \end{bmatrix},$$

$$\alpha_1 = 0, \alpha_2 = 0.8, \psi_1 = 1.6, \psi_2 = 0.4.$$

We first simulated normally distributed true factor scores (η_1 and η_2) with the exact means and variances, and then simulated item scores based on Equation 1 with normally distributed $\boldsymbol{\varepsilon}$. We then examined the relationship between sum scores (across three items) and the true latent variable scores, and between the regression factor scores and the true latent variable scores. Because invariance holds, if we take the sum score of the items for each person, the relationship between the sum score and η is the same across groups, as shown in Figure 1a, where the fitted lines between sum scores and the true latent variable scores are the same across groups.

Now, let's consider estimated factor scores. To obtain factor scores we need to first fit a confirmatory factor model to the data. Because invariance holds in this example, we can just fit the invariance model with equal loadings, intercepts, and unique variances across groups (i.e., strict invariance, Horn & McArdle, 1992).

Using the *lavaan* R package (Rosseel, 2012), we confirm that the invariance model fits well with the simulated data, with all parameter estimates being very close to the

values we used to generate the data. The details can be found in the online supplemental materials (https://github.com/marklhc/fsinv-supp/).

However, as shown in Figure 1b, using the regression method (implemented in *lavaan* but the algorithm is the same in *Mplus*), the factor scores obtained from the invariance model are not invariant, as the fitted lines with the factor scores, $\tilde{\eta}$, regressed on η differ across groups, in both the intercepts and the slopes. For example, for a participant in Group 1 with η = -0.5, the expected value of $\tilde{\eta}$ is -0.45; for a participant in Group 2 with the same level of η = -0.5, the expected value of $\tilde{\eta}$ is -0.1. Similar observations have been reported in Davoudzadeh et al. (2020). We will later prove that regression factor scores are not invariant, after the partial invariance example.

Example 2: Partial Invariance

In the second example, we make one modification to Group 2 such that $\boldsymbol{\nu}_2 = [0, 0, 0.5]^\top$, so Group 2 tends to have a higher score on item 3, when holding constant the latent variable of interest. So only partial invariance holds in the population. As shown in Figure 1c and as expected, the sum scores are not invariant, with the regression line for Group 2 uniformly shifted upward.

The noninvariance of item 3 can be accounted for by fitting a partial invariance model, where the intercept for item 3 is allowed unequal across groups. As shown in the supplemental materials, such a model fits the data well, and we can recover the latent means and variances used to simulate the data. However, while allowing item 3 to have different intercepts puts the latent variables on the same metric across groups, this is not preserved when computing factor scores based on the regression method, as shown in Figure 1d. Indeed, the fitted lines are the same as in Figure 1b, where they differ in both intercepts and slopes, meaning that the regression factor scores are neither metric invariant nor scalar invariant.

Proof That Regression Factor Scores Are Not Invariant

We will prove here that under a unidimensional factor model (Equation 1), regression factor scores are generally not measurement invariant, and the results generalize to multidimensional settings. Estimated factor scores, $\tilde{\eta}$, are weighted composites of the item scores (e.g., Devlieger et al., 2016). When considering the mean structure, the general form of $\tilde{\eta}$ is (e.g., Muthén, 2004)

$$\tilde{\eta}_g = \tilde{\alpha}_g + \mathbf{A}_g(\mathbf{y}_g - \boldsymbol{\mu}_{yg}), \tag{2}$$

where $\boldsymbol{\mu}_{yg} = \boldsymbol{\nu}_g + \boldsymbol{\lambda}_g\tilde{\alpha}_g$ is a vector of model-implied item group means and $\mathbf{A}$ is a $1 \times p$ factor scoring matrix applied to the centered data. Different choices of $\mathbf{A}$ correspond to different methods for computing factor scores. Note that here, $\tilde{\alpha}$ is the latent variable mean in the scoring stage, and does not need to correspond to the α values in the estimation stage of the measurement model, although most software by default sets $\tilde{\alpha}$ to the α value in the estimated measurement model.

Combining with Equation 1, we can express a measurement model of $\tilde{\eta}$ for the latent variable

$$\begin{aligned} \tilde{\eta}_g &= \tilde{\alpha}_g + \mathbf{A}_g(\boldsymbol{\nu}_g + \boldsymbol{\lambda}_g\eta_g + \boldsymbol{\varepsilon}_g - \boldsymbol{\nu}_g - \boldsymbol{\lambda}_g\tilde{\alpha}_g) \\ &= \underbrace{(1 - \mathbf{A}_g\boldsymbol{\lambda}_g)\tilde{\alpha}_g}_{\text{measurement intercept}} + \underbrace{\mathbf{A}_g\boldsymbol{\lambda}_g}_{\text{loading}}\eta_g + \underbrace{\mathbf{A}_g\boldsymbol{\varepsilon}_g}_{\text{measurement error}} \end{aligned} \tag{3}$$

Therefore, factor scores are metric invariant only when

$$\mathbf{A}_1\boldsymbol{\lambda}_1 = \mathbf{A}_2\boldsymbol{\lambda}_2, \tag{4}$$

scalar invariant when, in addition,

$$(1 - \mathbf{A}_1\boldsymbol{\lambda}_1)\tilde{\alpha}_1 = (1 - \mathbf{A}_2\boldsymbol{\lambda}_2)\tilde{\alpha}_2, \tag{5}$$

and strict invariant if, in addition,

$$\mathbf{A}_1\boldsymbol{\Theta}_1\mathbf{A}_1^\top = \mathbf{A}_2\boldsymbol{\Theta}_2\mathbf{A}_2^\top \tag{6}$$

For regression factor scores (Skrondal & Laake, 2001; Thomson, 1935),

$$\mathbf{A}_g = \tilde{\psi}_g \boldsymbol{\lambda}_g^\top (\tilde{\psi}_g \boldsymbol{\lambda}_g \boldsymbol{\lambda}_g^\top + \boldsymbol{\Theta}_g)^{-1}, \tag{7}$$

which depends on, $\tilde{\psi}$, the latent variance for scoring purposes. Note again that $\tilde{\psi}$ does not need to correspond to the ψ values in the estimation stage of the measurement model; however, because in most software (e.g., *lavaan* and *Mplus*), group-specific ψ and α from the estimated measurement model are used for factor score computations for Equation 2, regression factor scores are generally not measurement invariant unless the latent means and variances are constant across groups, *even when measurement invariance holds for items.* Thus, we see the results in Figure 1b.

On the other hand, for sum scores $Z_g = \mathbf{1}^\top \mathbf{y}_g$, we have

$$Z_g = \underbrace{\mathbf{1}^\top \boldsymbol{\nu}_g}_{\text{measurement intercept}} + \underbrace{(\mathbf{1}^\top \boldsymbol{\lambda}_g)}_{\text{loading}} \eta_g + \underbrace{\mathbf{1}^\top \boldsymbol{\varepsilon}_g}_{\text{measurement error}},$$

so one can see that sum scores are metric invariant when $\boldsymbol{\lambda}$ is equal across groups, scalar invariant when $\boldsymbol{\lambda}$ and $\boldsymbol{\nu}$ are equal across groups, and strict invariant when, additionally, $\boldsymbol{\Theta}$ is equal across groups.[1]

The case of Bartlett factor scores

On the other hand, factor scores based on the Bartlett method (Bartlett, 1937), which is available in most structural equation modeling programs, has

$$\mathbf{A}_g = (\boldsymbol{\lambda}_g^\top \boldsymbol{\Theta}_g^{-1} \boldsymbol{\lambda}_g)^{-1} \boldsymbol{\lambda}_g^\top \boldsymbol{\Theta}_g^{-1},$$

and so $\mathbf{A}_g \boldsymbol{\lambda}_g = 1$, meaning that such scores always have a loading of 1 and an intercept of 0 when treated as an indicator of the latent variable. Thus, factor scores based on the Bartlett method are generally at least scalar invariant.

[1] Even when the items are not invariant, sum scores can still be invariant when the sums of loadings, intercepts, and error variances and covariances are equal across groups. This can happen when one group has a higher intercept on one item but a lower intercept on another item than the other group, and is related to the "unity-weights invariance" condition discussed in Horn and McArdle (1992).

On the other hand, the scoring matrix in Equation 7 for regression factor scores can be written as

$$\mathbf{A}_g = (\tilde{\psi}_g^{-1} + \boldsymbol{\lambda}_g^\top \boldsymbol{\Theta}_g^{-1} \boldsymbol{\lambda}_g)^{-1} \boldsymbol{\lambda}_g^\top \boldsymbol{\Theta}_g^{-1},$$

using a variant of the Woodbury identity (Petersen & Pedersen, 2012, p. 18, Equation 158), which looks similar to the scoring matrix for Bartlett factor scores except that $\tilde{\psi}_g^{-1}$ is added inside the first pair of parentheses, so typically $\mathbf{A}_g \boldsymbol{\lambda}_g < 1$, which produces shrinkage for regression factor scores.[2].

Remedy: Using Common Latent Distributions When Computing Regression Factor Scores

From the previous discussion, one reason that regression factor scores are not invariant even when measurement invariance holds for items is that group-specific latent means and variances are used when computing factor scores. One modification to alleviate the problem is to use common latent means and latent variances for each group when computing regression factor scores. This ensures that when invariance holds for the items, $\mathbf{A}$ is constant. Figure 2a shows this for the simulated data in Example 1, where we use $\tilde{\psi}$ = 1 and $\tilde{\alpha} = 0.5$ for both groups when computing factor scores.[3] While such an option is not available in software like *lavaan* or *Mplus*, one can use software that supports matrix operations (e.g., R and SPSS) with Equation 3 and Equation 7 to directly compute the regression factor scores with the same latent means and variances across groups.

An interesting observation is that factor scores depend on the item loadings and

[2] We thank an anonymous reviewer for pointing this out. It can be shown that $\mathbf{A}_g \boldsymbol{\lambda}_g$ is also the theoretical reliability for regression factor scores.

[3] The chosen common mean and variance for η are used to rescale the latent variable for scoring purposes, and do not need to correspond to the scaling of the latent variable in the parameter estimation stage. Other mean and variance values give the same invariance properties for regression factor scores under full invariance of the items.

unique variances and covariances (as in the scoring matrix $\mathbf{A}$ and Equation 4), but not the item intercepts (as in Equation 5), as the intercepts are cancelled out in Equation 3. Because of this, when partial scalar invariance holds (i.e., where $\boldsymbol{\lambda}_1 = \boldsymbol{\lambda}_2$ but the intercepts are different for some items) *and* $\boldsymbol{\Theta}$ is equal across groups, regression factor scores based on the partial invariance model and using the same latent distributions are invariant. However, when only partial metric invariance holds, or when $\boldsymbol{\Theta}$ is different across groups, using common latent means and variances still results in noninvariant factor scores, because when $\boldsymbol{\lambda}$ are not equal across groups, the resulting loadings (and intercepts) of $\tilde{\eta}$ as an indicator of η are still different. Using similar data as in Example 2 but with λ_3 for Group 2 changed to 10 (intentionally chosen so that the difference is clearer in the graph), Figure 2b confirms that regression factor scores are still not invariant even when supplying the same latent means and variances, as the intercepts and slopes are different. Also, the degree of measurement error in the factor scores is much smaller in Group 2 because the scoring matrix is different across groups due to the scoring matrix $\mathbf{A}$ being a function of the loadings and playing a role in Equation 6.

Multiple-Indicator Multiple Cause (MIMIC)

An alternative to multiple-group analysis in modeling partial invariance is to use the MIMIC model. Here we limit our discussion to traditional MIMIC with continuous indicators. However, as recent methods such as moderated nonlinear factor analysis (MNLFA, Bauer & Curran, 2016) have MIMIC as the simplest case, our results should generalize to those methods, which have gained in popularity for yielding factor scores to be pooled in integrative data analysis.

To demonstrate, we fit the MIMIC model for the above-discussed simulated data for (a) Example 1 (invariance holds) and (b) Example 2 (partial invariance holds). For (a), we only include the effect of the grouping variable on the latent factor, whereas for (b), the grouping variable predicts both the latent factor and Item 3 as Item 3 is not scalar invariant. We used *Mplus* 8.8 and *lavaan* to fit the MIMIC models and obtain the

regression factor scores. As shown in Figure 3, regression factor scores based on the MIMIC model are invariant for neither (a) nor (b). Specifically, given the same level of the latent variable, factor scores for Group 2 are generally higher, as Group 2 has a larger latent mean. The Appendix gives more explanation, but the intuition is that in structural equation modeling, the MIMIC model is handled by treating the grouping variable as a latent variable with known means and variances, and when there are multiple latent variables, regression factor scores for a given latent variable generally are functions of all other latent variables, including the grouping variable. Thus, when computing scores for η, the scoring matrix is a function of the grouping variable, which violates invariance.[4] Similar observations have again been reported in Davoudzadeh et al. (2020).

Factor Scores for Item Response Models

While the above discussion focuses on factor scores on continuous indicators, mainly because the math is more tractable, it generalizes to scores based on item response models or other models for discrete items. For example, let $f(\mathbf{y} \mid \eta, \boldsymbol{\omega})$ be the likelihood function under any measurement model linking item scores $\mathbf{y}$ and latent variable η with measurement parameters $\boldsymbol{\omega}$. The commonly used scores based on the expected a posteriori (EAP) method can be obtained as the mean of the following posterior distribution of η:

$$\frac{f(\mathbf{y} \mid \eta, \boldsymbol{\omega})\pi(\eta)}{\int f(\mathbf{y} \mid h, \boldsymbol{\omega})\pi(h)\, dh},$$

where $\pi(\cdot)$ is the prior distribution for the latent variable, and is typically set to the same distribution of the latent variables used in the estimation stage (i.e., normal distribution with specific means and variances). When the measurement model is chosen as the linear factor model discussed above, the EAP method is the same as the factor scores based on

[4] As pointed out by an anonymous reviewer, there are alternative ways to estimate factor scores than the one used in *Mplus* and *lavaan* at the time of this writing. For example, Skrondal and Laake (2001) discussed methods for separately estimating factor scores for each individual latent variable or each block of latent variables, which may result in a scoring matrix that is not a function of the grouping variable.

the regression method (see Thissen & Thissen-Roe, 2020). Also, the maximum a posteriori (MAP) method, which is based on the posterior mode, gives the same factor scores as the posterior distribution is symmetric under the linear factor model (Thissen & Thissen-Roe, 2020). On the other hand, the maximum likelihood estimates of factor scores do not depend on priors and are obtained by maximizing the likelihood function, which reduces to Bartlett factor scores under the linear factor model (Thissen & Thissen-Roe, 2020).

Therefore, like regression factor scores, the EAP and MAP scores are also noninvariant even under a multiple-group measurement invariance model, when group-specific latent means and variances are used for scoring. Figure 4a shows an example similar to Example 1, but with nine dichotimized items. Specifically, we used the item factor model specification (Wirth & Edwards, 2007) by assuming that each binary item has an underlying normal variate. All loadings are 0.8 and all intercepts and thresholds are 0. We used the *mirt* R package (Chalmers, 2012) to estimate a multiple-group 2-parameter logistic item response model, with the threshold and discrimination parameters held invariant across groups (i.e., no differential item functioning [DIF]). The figure shows that the relation between the EAP scores and the latent variables (which are known because we use simulated data) is different across the two groups.

On the other hand, EAP scores are invariant when the same latent distributions are used for scoring for both groups, when the items are invariant (i.e., no DIF), as shown in Figure 4b. Such an option is available in the `fscores()` function of *mirt* R package by specifying the `mean` and `cov` arguments.

We have shown that, even under full invariance of items, EAP scores are not invariant when computed using group-specific latent distributions, and it is easy to infer and show that EAP scores with group-specific latent distributions are also not invariant under partial invariance of items. Here we will show that, like regression factor scores, the EAP scores are not invariant under a partial invariance model (i.e., when DIF is modeled) even when supplying the same latent mean and variance. Given that observed scores are

not linear functions of the latent variables in item response models, it is harder to visualize the noninvariance of EAP scores using just one simulated data set; instead, we conducted a small Monte Carlo simulation of partial invariance for binary items. Specifically, three out of nine items are noninvariant; for Group 1, all loadings are 0.8 and all intercepts are 0, whereas the noninvariant loadings are 0.4 and the noninvariant intercepts are 0.5 for Group 2. All thresholds are 0. We performed 5,000 replications, each with 100 observations per group. We simulated the latent scores to be exactly the same for both groups and across replications to minimize sampling errors. Therefore, if EAP scores are invariant, they should have identical distributions across groups for every level of the latent variable, η. Figure 5 shows the contrary, as the scores are higher for Group 1 for a higher level of η, and higher for Group 2 for a lower level of η.

Discussion

Since the development of factor analysis, different factor-scoring mechanisms have been proposed to score the latent variables. Among the most popular are the regression method and the Bartlett method. The regression method, which produces shrinkage scores according to the reliability of the indicators, also generalizes to the commonly used EAP and MAP scores in item response models. While such scores are primarily used for assessment purposes in estimating individuals' standing on some latent trait (mostly the g factor) in the early days, over the years, researchers have been interested in using those scores in statistical analyses. Especially after the development of structural equation modeling, there has been interest in using either sum scores or factor scores as proxies of abstract latent variables so that models can be simplified (e.g., Skrondal & Laake, 2001). While earlier applications of factor scores concern analyses of a single group, more recent applications—including data harmonization across studies and waves in longitudinal analyses—concern analyses of multiple samples. The goal of such analyses is to obtain scores that are comparable across samples. Unfortunately, as highlighted earlier, there seems to be a conception that in a model where the latent variables are scaled to the same

metric, the corresponding factor scores will be on the same metric as well (e.g., Curran & Hussong, 2009; McNeish, 2022; Steenkamp & Maydeu-Olivares, 2021). To our knowledge, there has been no systematic discussion on whether factor scores are on the same metric, even when partial invariance is accounted for.

As shown in the above examples, like unweighted sum scores, regression factor scores—which are weighted sums of items scores—are not invariant when some items are noninvariant. Such a conclusion generalizes to EAP or MAP scores that are popular when using item response models. Therefore, obtaining these factor scores does *not* result in variables that are on the same metric across samples or time, even when partial invariance (or DIF) is accounted for. This is in contrast to some recommended practices of data harmonization or score linking (e.g., Curran & Hussong, 2009; McArdle et al., 2009). Note that the condition of partial invariance is the same as in linkage of measurement based on common items—where the common items are assumed invariant and the unique items are assumed noninvariant (as they are different items), and thus our conclusions would generalize to such situations. Perhaps more importantly, when computed based on a multiple-group analysis with group-specific latent means and variances, which are usually the default in popular software packages (e.g., *Mplus*, *lavaan*, *mirt*), shrinkage scores are even noninvariant when full invariance of items hold. Therefore, we strongly caution researchers when inferring differences across groups/time with such scores.

One thing to clarify is that, while shrinkage factor scores are not measurement invariant, these scores may preserve some properties of the latent distributions. For example, looking at Equation 3, one can see that regression scores based on group-specific latent distributions preserve the latent mean differences. On the other hand, using shrinkage scores as the response variable in regression results in biased regression coefficients, which is well documented in the literature (e.g., Skrondal & Laake, 2001). Shrinkage scores also generally resulted in biased standardized coefficients as they tend to

have smaller variances than the true latent variables (Devlieger et al., 2016).[5] In an example in the supplemental materials, we showed that because regression factor scores are generally not metric invariant when using group-specific latent means and variances, their use can lead to spurious interaction effects. Furthermore, using scores that are measurement invariant does not necessarily lead to unbiased parameter estimation; for example, while we have shown that scores based on the Bartlett method are invariant, their use as predictors in regression also results in biased regression coefficients (Skrondal & Laake, 2001), because measurement error in them is not accounted for.

The take-home message is that measurement models are used to calibrate the latent variables on the same metric, not the factor scores. To obtain valid inferences involving latent variables, instead of using factor scores, one can directly specify a hypothesized model concerning latent variables on top of the partial invariance models. Such a practice is routine in structural equation modeling, where researchers jointly model the structural and the measurement parts of the model, such as in second-order growth analysis (e.g., Ferrer et al., 2008). The joint modeling approach not only calibrates the latent variables on the same metric, but also accounts for the measurement errors in the indicators. On the other hand, the appeal of using factor scores is that they allow for a divide-and-conquer approach so that researchers can focus their energy on the measurement part first, and then deal with a structural model that has fewer variables and parameters. Fortunately, there have been several promising approaches that lead to valid inferences with factor scores, by accounting for the relation between factor scores and the latent variables. These include factor score regression (e.g., Croon, 2002; Devlieger et al., 2016), the

[5] For example, in Example 1 with full invariance, the pooled standard deviation (SD) of the latent variables across the two groups is $\sqrt{(1.6 + 0.4)/2} = 1$, and the standardized mean difference is 0.8. For regression factor scores, if we use group-specific latent means and variances, the unstandardized mean difference is 0.808, and the pooled SD is 0.93, so the standardized mean difference is 0.87. If we use common latent means and variances, the unstandardized mean difference is 0.64, and the pooled SD is 0.87, so the standardized mean difference is 0.74.

structural-after-measurement approach (Rosseel & Loh, 2022), and two-stage path-analysis (Lai et al., 2023; Lai & Hsiao, 2022). We strongly encourage researchers to adopt these approaches when utilizing factor scores for analysis.

References

Bartlett, M. S. (1937). The statistical conception of mental scores. *British Journal of Psychology. General Section*, *28*(1), 97–104. https://doi.org/10.1111/j.2044-8295.1937.tb00863.x

Bauer, D. J., & Curran, P. J. (2016). The discrepancy between measurement and modeling in longitudinal data analysis. In J. R. Harring, L. M. Stapleton, & S. N. Beretvas (Eds.), *Advances in multilevel modeling for educational research: Addressing practical issues found in realworld applications* (pp. 3–38). Information Age.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456–466. https://doi.org/10.1037/0033-2909.105.3.456

Chalmers, R. P. (2012). **mirt**: A Multidimensional Item Response Theory Package for the*R*Environment. *Journal of Statistical Software*, *48*(6). https://doi.org/10.18637/jss.v048.i06

Croon, M. A. (2002). Using predicted latent scores in general latent structure models. In G. A. Marcoulides & I. Moustaki (Eds.), *Latent variable and latent structure models* (pp. 195–224). Lawrence Erlbaum.

Curran, P. J., Cole, V., Bauer, D. J., Hussong, A. M., & Gottfredson, N. (2016). Improving factor score estimation through the use of observed background characteristics. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(6), 827–844. https://doi.org/10.1080/10705511.2016.1220839

Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods*, *14*(2), 81–100. https://doi.org/10.1037/a0015914

Curran, P. J., McGinley, J. S., Bauer, D. J., Hussong, A. M., Burns, A., Chassin, L., Sher, K., & Zucker, R. (2014). A moderated nonlinear factor model for the development of

commensurate measures in integrative data analysis. *Multivariate Behavioral Research*, *49*(3), 214–231. https://doi.org/10.1080/00273171.2014.889594

Davoudzadeh, P., Grimm, K. J., Widaman, K. F., Desmarais, S. L., Tueller, S., Rodgers, D., & Van Dorn, R. A. (2020). Estimation of latent variable scores with multiple group item response models: Implications for integrative data analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(6), 931–941. https://doi.org/10.1080/10705511.2020.1724113

Devlieger, I., Mayer, A., & Rosseel, Y. (2016). Hypothesis testing using factor score regression: A comparison of four methods. *Educational and Psychological Measurement*, *76*(5), 741–770. https://doi.org/10.1177/0013164415607618

Ferrer, E., Balluerka, N., & Widaman, K. F. (2008). Factorial invariance and the specification of second-order latent growth models. *Methodology*, *4*(1), 2236. https://doi.org/10.1027/1614-2241.4.1.22

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*(3-4), 117–144. https://doi.org/10.1080/03610739208253916

Lai, M. H. C., & Hsiao, Y.-Y. (2022). Two-stage path analysis with definition variables: An alternative framework to account for measurement error. *Psychological Methods*, *27*(4), 568–588. https://doi.org/10.1037/met0000410

Lai, M. H. C., Richardson, G. B., & Wa Mak, H. (2019). Quantifying the impact of partial measurement invariance in diagnostic research: An application to addiction research. *Addictive Behaviors*, *94*, 50–56. https://doi.org/10.1016/j.addbeh.2018.11.029

Lai, M. H. C., & Tse, W. W.-Y. (2023). *Are factor scores measurement invariant?* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/uzrak

Lai, M. H. C., Tse, W. W.-Y., Zhang, G., Li, Y., & Hsiao, Y.-Y. (2023). Correcting for Unreliability and Partial Invariance: A Two-Stage Path Analysis Approach. *Structural Equation Modeling: A Multidisciplinary Journal*, *30*(2), 258–271.

https://doi.org/10.1080/10705511.2022.2125397

Luningham, J. M., McArtor, D. B., Hendriks, A. M., Beijsterveldt, C. E. M. van, Lichtenstein, P., Lundstrm, S., Larsson, H., Bartels, M., Boomsma, D. I., & Lubke, G. H. (2019). Data integration methods for phenotype harmonization in multi-cohort genome-wide association studies with behavioral outcomes. *Frontiers in Genetics*, *10*. https://doi.org/10.3389/fgene.2019.01227

Luong, R., & Flake, J. K. (2022). Measurement invariance testing using confirmatory factor analysis and alignment optimization: A tutorial for transparent analysis planning and reporting. *Psychological Methods*. https://doi.org/10.1037/met0000441

MacDonald, G., & Park, Y. (2022). Associations of attachment avoidance and anxiety with life satisfaction, satisfaction with singlehood, and desire for a romantic partner. *Personal Relationships*, *29*(1), 163–176. https://doi.org/10.1111/pere.12416

McArdle, J. J., Grimm, K. J., Hamagami, F., Bowles, R. P., & Meredith, W. (2009). Modeling life-span growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. *Psychological Methods*, *14*(2), 126–149. https://doi.org/10.1037/a0015857

McNeish, D. (2022). Psychometric properties of sum scores and factor scores differ even when their correlation is 0.98: A response to Widaman and Revelle. *Behavior Research Methods*, *55*(8), 4269–4290. https://doi.org/10.3758/s13428-022-02016-x

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika*, *29*(2), 177–185. https://doi.org/10.1007/BF02289699

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge. https://www.taylorfrancis.com/books/9780203821961

Millsap, R. E., & Kwok, O.-M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods*, *9*(1), 93–115. https://doi.org/10.1037/1082-989X.9.1.93

Muthén, B. (2004). *Mplus technical appendices*.

https://www.statmodel.com/download/techappen.pdf

Petersen, K. B., & Pedersen, M. S. (2012). *The matrix cookbook*. Technical University of Denmark. http://www2.compute.dtu.dk/pubdb/pubs/3274-full.html

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Rosseel, Y., & Loh, W. W. (2022). A structural after measurement approach to structural equation modeling. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000503

Shi, D., Song, H., & Lewis, M. D. (2019). The impact of partial factorial invariance on cross-group comparisons. *Assessment*, *26*(7), 1217–1233. https://doi.org/10.1177/1073191117711020

Skrondal, A., & Laake, P. (2001). Regression among factor scores. *Psychometrika*, *66*(4), 563–575. https://doi.org/10.1007/BF02296196

Steenkamp, J.-B. E. M., & Maydeu-Olivares, A. (2021). An updated paradigm for evaluating measurement invariance incorporating common method variance and its assessment. *Journal of the Academy of Marketing Science*, *49*(1), 5–29. https://doi.org/10.1007/s11747-020-00745-z

Thissen, D., & Thissen-Roe, A. (2020). Factor score estimation from the perspective of item response theory. In M. Wiberg, D. Molenaar, J. González, U. Böckenholt, & J.-S. Kim (Eds.), *Quantitative psychology: 84th annual meeting of the Psychometric Society, Santiago, Chile, 2019* (Vol. 322, pp. 171–184). Springer International Publishing. https://doi.org/10.1007/978-3-030-43469-4_14

Thomson, G. H. (1935). Definition and measurement of general intelligence. *Nature*, *135*(3413), 509–509. https://doi.org/10.1038/135509b0

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances*

from alcohol and substance abuse research. (pp. 281–324). American Psychological Association. https://doi.org/10.1037/10222-009

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*(1), 58–79. https://doi.org/10.1037/1082-989X.12.1.58

Zhao, X., Coxe, S., Sibley, M. H., Zulauf-McCurdy, C., & Pettit, J. W. (2022). Harmonizing Depression Measures Across Studies: a Tutorial for Data Harmonization. *Prevention Science*. https://doi.org/10.1007/s11121-022-01381-5

Appendix

MIMIC factor scores

In this appendix, we will show that regression factor scores are not invariant when the latent variable is predicted by a grouping variable (or a covariate). In *Mplus* and *lavaan*, under MIMIC, the grouping variable G is represented as a latent variable. Therefore, for a unidimensional factor model with respect to one grouping variable, Ψ is of dimension 2×2. Assume that the model is identified by letting the conditional variance of η be 1,

$$\Psi = \begin{bmatrix} 1 & \\ 0 & V(G) \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 0 & \beta \\ 0 & 0 \end{bmatrix}, \Lambda = \begin{bmatrix} \boldsymbol{\lambda} & 0 \\ 0 & 1 \end{bmatrix},$$

$$\boldsymbol{\Theta} = \begin{bmatrix} \boldsymbol{\Theta}_y & 0 \\ 0 & 0 \end{bmatrix},$$

where β is the direct path from G to η, $V(G)$ is the variance of the grouping variable, $\boldsymbol{\lambda}$ is the loadings of $\mathbf{y}$ on η, and $\boldsymbol{\Theta}_y$ is the unique covariance matrix of $\mathbf{y}$.

With MIMIC, the regression factor scores $(\tilde{\eta}, \tilde{G})$ for (η, G) are computed with the scoring matrix

$$\mathbf{A} = \Sigma_{\text{lv}} \boldsymbol{\Lambda}^\top [\boldsymbol{\Lambda} \Sigma_{\text{lv}} \boldsymbol{\Lambda}^\top + \boldsymbol{\Theta}]^{-1}, \tag{8}$$

where $\Sigma_{\text{lv}} = (\mathbf{I} - \mathbf{B})^{-1} \Psi [(\mathbf{I} - \mathbf{B})^{-1}]^\top$ is the implied covariance matrix of (η, G). $\mathbf{A}$ can be partitioned as

$$\begin{array}{c} \\ \tilde{\eta} \\ \tilde{G} \end{array} \begin{array}{c} \begin{array}{cc} \mathbf{y} & G \end{array} \\ \begin{bmatrix} \mathbf{a}_1 & a_2 \\ \mathbf{a}_3 & a_4 \end{bmatrix} \end{array},$$

where $\mathbf{a}_1$ is the row vector of scoring weights for $\tilde{\eta}$ from the items, and a_2 is the weight for computing $\tilde{\eta}$ from G. In other words, factor scores for $\tilde{\eta}$ are computed as $\mathbf{a}_1 \mathbf{y} + a_2 G$. When

the items are invariant, a nonzero a_2 would mean that the factor scores are not invariant, because $\tilde{\eta}$ is a function also of G. Note that the factor scores $\tilde{G}$ for G are also computed, but generally one gets $\tilde{G} = G$ and they are not of interest.

Because a_2 is the last element from the first row of $\mathbf{A}$, from Equation 8, we only need to consider the product of the first row of $\Sigma_{\text{lv}}\boldsymbol{\Lambda}^\top$ and the last column of $[\boldsymbol{\Lambda}\Sigma_{\text{lv}}\boldsymbol{\Lambda}^\top + \boldsymbol{\Theta}]^{-1}$, the latter being the inverse of the implied covariance of $(\mathbf{y}, G)$. The first row of $\Sigma_{\text{lv}}\boldsymbol{\Lambda}^\top$ is $[\boldsymbol{\lambda}^\top|\beta]$. Let $V(\mathbf{y}) = V(\eta)\boldsymbol{\lambda}\boldsymbol{\lambda}^\top + \boldsymbol{\Theta}_y$ be the implied covariance matrix of $\mathbf{y}$. The implied covariance matrix of $(\mathbf{y}, G)$ is

$$\boldsymbol{\Lambda}\Sigma_{\text{lv}}\boldsymbol{\Lambda}^\top + \boldsymbol{\Theta} = \begin{bmatrix} V(\mathbf{y}) & \beta V(G)\boldsymbol{\lambda} \\ \beta V(G)\boldsymbol{\lambda}^\top & V(G) \end{bmatrix}.$$

Because the above covariance matrix is positive definite, its inverse cannot have a column of all zeros. Given that a_2 is computed as the matrix multiplication of the first row of $\Sigma_{\text{lv}}\boldsymbol{\Lambda}^\top$ and the second column of the inverse of the implied covariance matrix, and that $\boldsymbol{\lambda} \neq \mathbf{0}$, $a_2 \neq 0$ unless $\beta = 0$ (in which case the implied covariance matrix and its inverse are block diagonal). Thus, regression factor scores under MIMIC are not invariant even when invariance holds for the items.

Similar steps can be used to show that regression factor scores are not invariant under MIMIC with partial invariance, where a direct path is allowed from G to one or more items in $\mathbf{y}$. In structural equation model representation of MIMIC with $\tilde{p}$ noninvariant items, the noninvariant items are also treated as latent variables so that $\boldsymbol{\Lambda}$, $\boldsymbol{\Psi}$, and $\mathbf{B}$ have expanded dimensions.

Figure 1

Observed scores against true latent factor under invariance (a and b) or partial invariance (c and d).

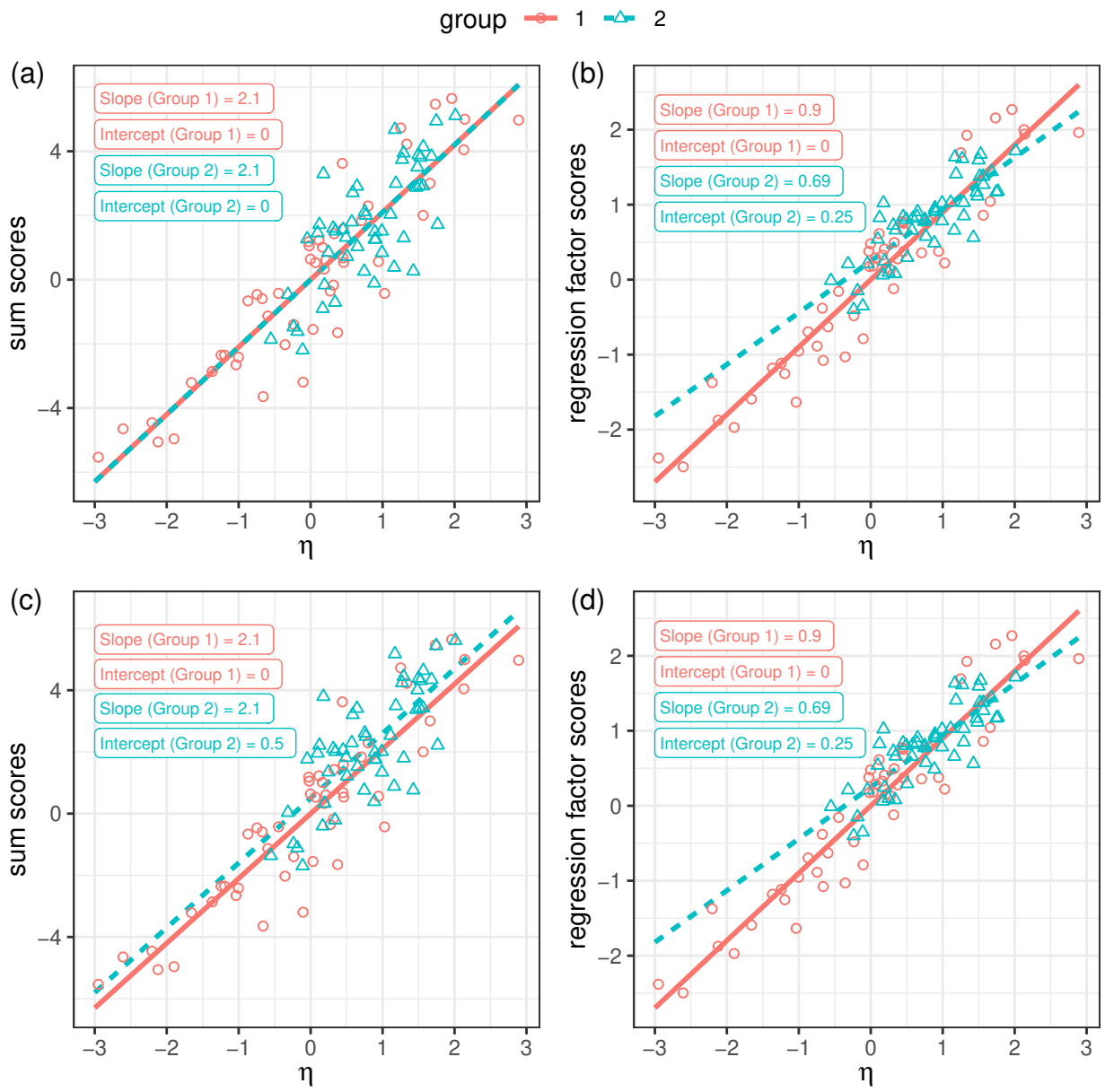

Figure 2

Regression factor scores against true latent factor under (a) full invariance and (b) partial metric invariance when using the same latent distribution.

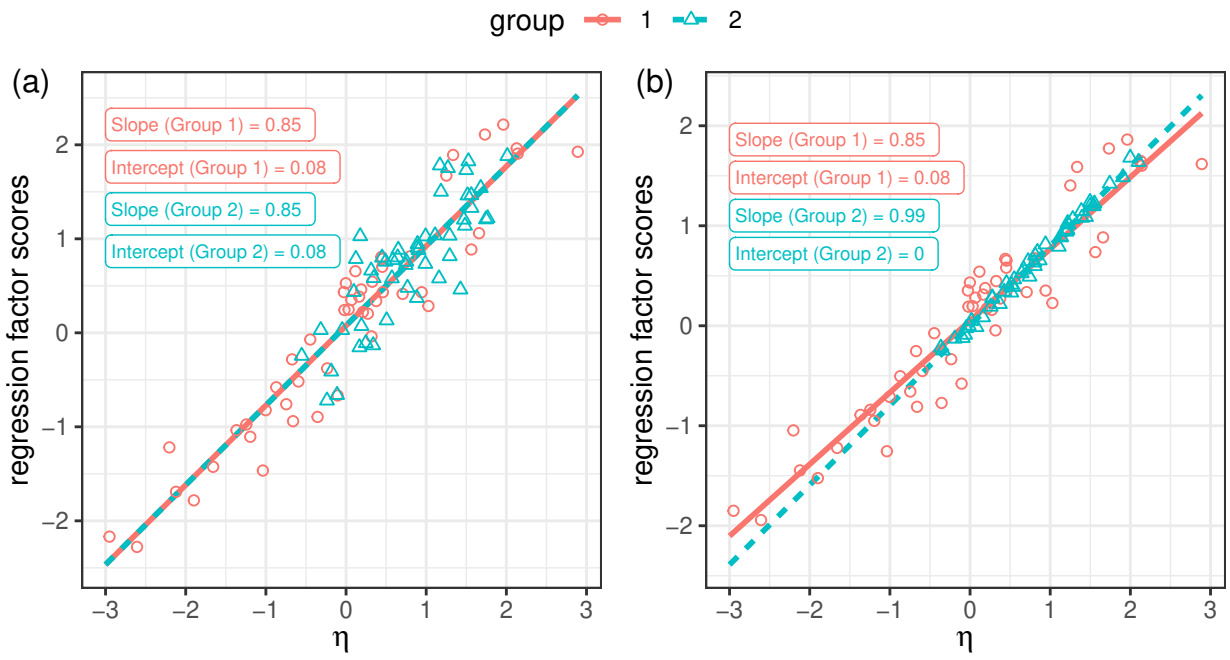

Figure 3

Regression factor scores against true latent factor under MIMIC with (a) invariance and (b) partial scalar invariance.

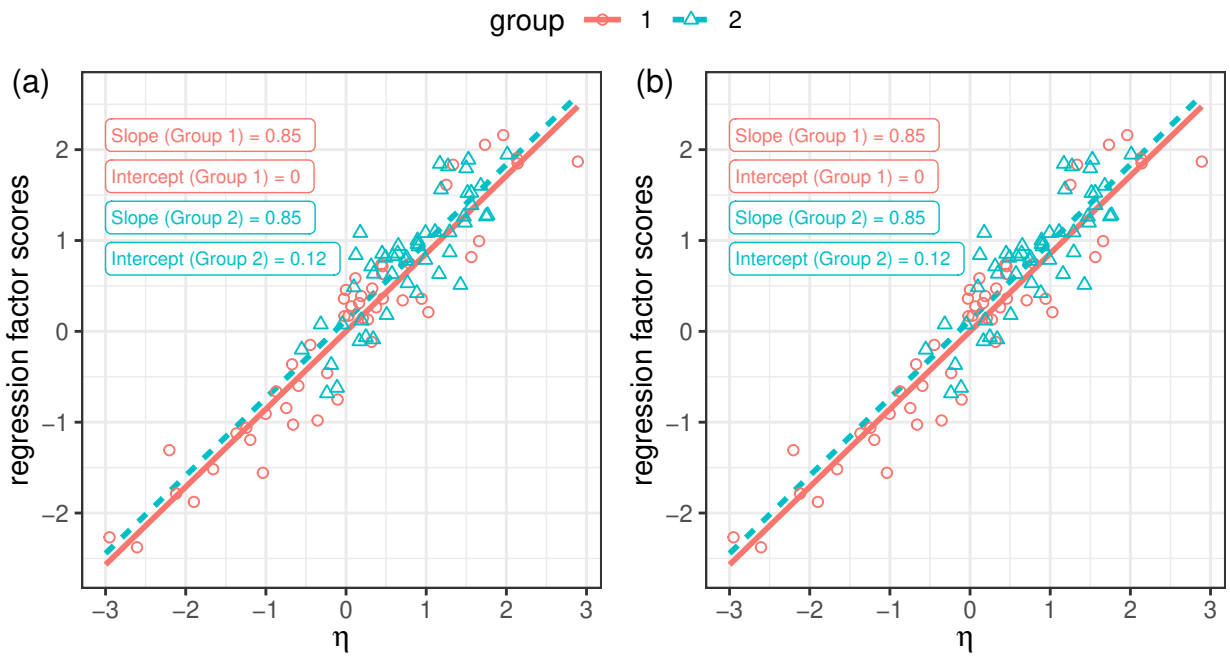

Figure 4

Expected a posteriori (EAP) scores against true latent factor under full invariance, with (a) group-specific and (b) equal means and variances. The fitted curves are obtained using natural cubic splines.

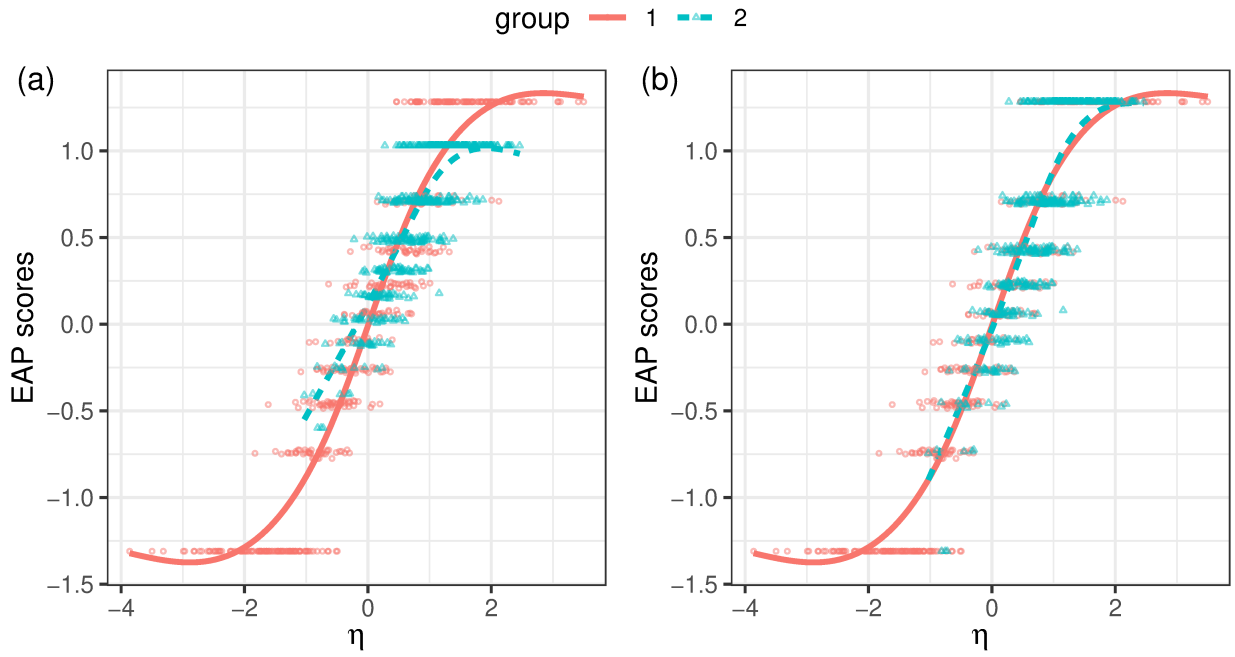

Figure 5

Expected a posteriori (EAP) scores against true latent factor under partial invariance across 5,000 replications. The vertical bars show the 10th and 90th percentiles, and the lines show the median across replications.