Adjusting for partial invariance in latent parameter estimation: Comparing forward

specification search and approximate invariance methods

Mark H. C. Lai[1], Yuanfang Liu[2], & Winnie Wing-Yee Tse[1]

[1] Department of Psychology, University of Southern California

[2] School of Education, University of Cincinnati

Author Note

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

Winnie Wing-Yee Tse https://orcid.org/0000-0001-5175-6754

Correspondence concerning this article should be addressed to Mark H. C. Lai, 3620 South McClintock Ave., Los Angeles, CA 90089-1061. E-mail: hokchiol@usc.edu

Abstract

Measurement invariance is the condition that an instrument measures a target construct in the same way across subgroups, settings, and time. In psychological measurement, usually only partial, but not full, invariance is achieved, which potentially biases subsequent parameter estimations and statistical inferences. Although existing literature shows that a correctly specified partial invariance model can remove such biases, it ignores the model uncertainty in the specification search step: flagging the wrong items may lead to additional bias and variability in subsequent inferences. On the other hand, several new approaches, including Bayesian approximate invariance and alignment optimization methods, have been proposed; these methods use an *approximate* invariance model to adjust for partial measurement invariance without the need to directly identify noninvariant items. However, there has been limited research on these methods in situations with a small number of groups. In this paper, we conducted three systematic simulation studies to compare five methods for adjusting partial invariance. While specification search performed reasonably well when the proportion of noninvariant parameters was no more than $1/3$, alignment optimization overall performed best across conditions in terms of efficiency of parameter estimates, confidence interval coverage, and Type I error rates. In addition, the Bayesian version of alignment optimization performed best for estimating latent means and variances in small-sample and low-reliability conditions. We thus recommend the use of the alignment optimization methods for adjusting partial invariance when comparing latent constructs across a few groups.

Keywords: measurement invariance, partial invariance, approximate invariance, item bias, specification search, alignment optimization

Word count: 8,444

Adjusting for partial invariance in latent parameter estimation: Comparing forward

specification search and approximate invariance methods

Measurement is the foundation of quantitative science. It is especially crucial for

psychological and behavioral sciences, as research results are highly susceptible to random

and systematic measurement errors in the instruments used. As a result, measurement

invariance, the condition that an instrument measures one or more constructs in the same

way across groups, has been a major research focus. For example, a quick search on the

PsycINFO database found 826 articles with the phrase "measurement invariance" or

"measurement equivalence" in the title or in the abstract, in just years 2018 and 2019.

While a main goal of evaluating measurement invariance is to detect the presence of

problematic items so that those items can be replaced or improved, establishing invariance

is also important because statistical inferences based on observed scores may be

compromised when invariance is violated for one or more items (Horn & McArdle, 1992).

The current paper thus focuses on methods for obtaining valid inferences when

measurement invariance does not hold.

Methodological scholars have shown that when one or more items in an instrument

show systematic bias across groups, a condition known as *partial invariance*, parameter

estimates and inferences of group comparisons on the latent construct will likely be biased

(Meredith, 1993). One way to correct for the bias is to use a partial invariance model,

where measurement parameters of the biased items are estimated without constraints

across groups. Although previous simulation studies found that using a partial invariance

model resulted in valid estimates and inferences of the latent parameters (e.g., Guenole &

Brown, 2014; Hsiao & Lai, 2018; Shi, Song, & Lewis, 2017), these studies mostly presumed

that researchers already knew which item(s) were invariant and could thus specify the

correct model. In practice, as pointed out in the review by Schmitt and Kuljanin (2008),

researchers usually engage in sequential *specification search*—iteratively freeing invariance

constraints—to identify biased items and form a partial invariance model. However, due to sampling error, the specification search step may not identify the correct partial invariance model (MacCallum et al., 1992), resulting in *model uncertainty*. As a result, using a partial invariance model based on a specification search may lead to less optimal estimates and inferences than what previous studies suggested.

As a simple heuristic example to illustrate model uncertainty, consider a scenario with a construct measured by four items in two groups, where item 4 is noninvariant, and the true latent mean difference, δ, is 0. Given a moderate sample size, using specification search has an 80% chance of getting the correct partial invariance model in the sample (i.e., with item 4 correctly identified as noninvariant), in which case the expected estimate of δ is 0. However, the specification search has a 20% chance of getting a wrong partial invariance model where item 4 is identified as invariant, which leads to biased δ estimates, with an expected estimate of, say, 0.2. Therefore, δ is unbiasedly estimated only when the correct partial invariance model is selected. When incorporating the model uncertainty, using specification search thus results in a bias of (.80) (0) + (.20) (.20) - 0 = .04 when estimating δ; however, when evaluating the performance of the specification method, previous studies have generally ignored model uncertainty.

On the other hand, when the main goal of research is to obtain valid statistical inferences instead of detecting noninvariant items, several methods based on the concept of *approximate invariance*—holding loadings and intercepts to be approximately, but not exactly, equal—have been proposed. These include the *alignment optimization (AO)* (Asparouhov & Muthén, 2014) and the *Bayesian approximate invariance with small variance priors* methods (Muthén & Asparouhov, 2013). A potential advantage of them is they do not require an iterative search process, but instead rely on an approximate invariance model that do not impose exact invariance assumptions for any items (i.e., they do not force any loadings or intercepts to be exactly equal across groups).[1] Previous

_____

[1] As pointed out by the Associate Editor, specification search assumes that *exact invariance* is achievable,

research (e.g., Kim et al., 2017; Marsh et al., 2018) has found that AO performed well in recovering latent parameters when the number of groups is large. However, given that a vast majority of measurement invariance research focused on two or three groups (Putnick & Bornstein, 2016), a critical question is whether these newer methods are effective for adjusting partial invariance in few groups, as compared to the conventional specification search method, which was developed in context with a few groups.

In the current paper, we present results from three Monte Carlo simulation studies comparing the accuracy of point estimates and coverage rates of interval estimates from specification search and several AO and approximate invariance methods. Unlike previous studies, the simulation studies take into account the model uncertainty in the specification step. Study 1 focuses on comparing various approaches in recovering the latent means with two groups, while Study 2 aims to replicate the findings of Study 1 with four groups. Study 3 focuses on latent regression coefficients of a latent predictor on an observed outcome. Below we first introduce the model notations and definitions of measurement invariance under the common factor model framework. We then describe each method to be compared in more detail and reviewed previous research findings, before transitioning to the design of the studies in the current paper.

Measurement and Factorial Invariance

Consider a scale of p items measuring a latent construct η. Let y_{ij} ($i = 1, 2, \ldots, N$; $j = 1, 2, \ldots, p$) be the score of the ith person on the jth item. A measurement model links

---

while approximate invariance methods makes a weaker assumption of approximate invariance. Therefore, when exact invariance is of interest, switching to an approximate invariance assumption may not be theoretically sound. In practice, however, we think the line between exact and approximate invariance is usually blurred: even though the likelihood ratio test in conventional invariance evaluation tests the exact invariance hypothesis, the test is usually rejected (e.g., Byrne et al., 1989; Schmitt & Kuljanin, 2008; Shi, Song, & Lewis, 2017), and researchers instead rely on alternative fit indices such as change in CFI and RMSEA (Chen, 2007; Cheung & Rensvold, 2002), which only indicate whether the invariance constraints hold approximately given the data, and thus is analogous to approximate invariance. We encourage researchers to carefully consider their research questions (e.g., testing exact invariance vs. obtaining valid mean comparisons) when selecting procedures for evaluating invariance.

y and η stochastically by specifying the conditional distribution $P(y_{ij}|\eta_i;\boldsymbol{\vartheta})$ with parameters $\boldsymbol{\vartheta}$. Formally, measurement invariance holds when the conditional distribution of the observed items is the same across subgroups (Mellenbergh, 1989; Meredith, 1993), such that for every subgroup k $(k = 1, 2, \ldots, K)$, like gender and ethnicity,

$$P(y_{ijk}|\eta_{ik}) \text{ does not depend on } k.$$

In other words, measurement invariance means that any two persons with the same η score have the same propensity to endorse any value on all items in the scale. Assuming the correct measurement model, measurement invariance holds when $\boldsymbol{\vartheta}$ is the same across groups.

With continuous items, the common factor model (Thurstone, 1947) is usually used, represented as

$$y_{ij} = \nu_j + \lambda_j \eta_i + \varepsilon_{ij}, \tag{1}$$

where ν_j is the measurement intercept, λ_j is the factor loading, and ε_{ij} is the unique factor. It is commonly assumed that ε is normally distributed with constant variance θ_j, so that y_{ij} is also normally distributed conditioned on η_i. In addition, the local independence assumption is usually applied such that, when conditioned on η_i, $\mathrm{Cov}(y_{ij}, y_{ij'}|\eta_i) = 0$ for $j \neq j'$, implying that the covariance matrix of εs is diagonal. When there are K groups, the model can be represented as

$$y_{ijk} = \nu_{jk} + \lambda_{jk} \eta_{ik} + \varepsilon_{ijk}. \tag{2}$$

When a common factor model holds, measurement invariance requires that the measurement parameters, meaning ν_j, λ_j, and θ_j for the model in (1), to be the same across groups, a condition called *factorial invariance* (Meredith, 1993). A common framework for evaluating factorial invariance includes four stages (e.g., Vandenberg & Lance, 2000): (a) configural invariance, which requires the configuration of the factor

loadings to be the same across groups (Horn & McArdle, 1992), meaning that each group should have the same number of latent constructs and each latent construct is measured by the same set of items across groups; (b) metric invariance, which requires, in addition to configural invariance, that the factor loadings are equal (i.e., $\lambda_{jk} = \lambda_j$ for all js and ks) to ensure that the difference between two values in the latent construct is comparable across groups; (c) scalar invariance, which requires, in addition to metric invariance, that the measurement intercepts to be the same (i.e., $\nu_{jk} = \nu_j$ for all js and ks) to ensure that the latent construct has the same zero point across groups; (d) strict invariance, or strict factorial invariance, which requires all measurement parameters (ν_j, λ_j, and θ_j for all js) to be equal across groups.

Specification Search

When there is evidence for violation of a stage of factorial invariance, researchers commonly conduct a specification search to identify biased items (Schmitt & Kuljanin, 2008). In practice, researchers commonly go through the four factorial invariance stages in a sequential order (e.g., Kline, 2016), and identify biased items within a stage if the equality constraints for that stage are rejected. For example, one first conducts a likelihood ratio test (LRT) comparing the configural invariance and the metric invariance models. If the LRT is statistically significant (usually at .05 level), indicating at least one of the $p \times K$ loadings is different from others, a specification search is conducted by consulting modification indices (MI; Sörbom, 1989), and the loading with the largest MI will be freed. We referred this process as the forward specification search (FS) in this paper.[2] [3] In the

————

[2] As one reviewer pointed out, nonsequential specification search methods have also been proposed, including the factor-ratio test (Cheung & Lau, 2012) and the Bayesian SEM approach (Shi, Song, Liao, et al., 2017). In this paper, we chose to focus on FS given that it is commonly done in the literature (e.g., Schmitt & Kuljanin, 2008).

[3] The approach was labelled as the backward approach in Jung and Yoon (2016). We chose to follow Marsh et al. (2018) to call it the forward approach as it is more consistent with the regression literature from a more restrictive model to a less restrictive model.

modified model, MIs will be recomputed, and the process of freeing loadings and recomputing MIs is repeated until none of the MIs of the loadings is larger than a prespecified cutoff, usually 3.84, which is the 95th percentile of a χ^2 distribution with one degree of freedom. Moving on to the next stage, this partial metric invariance model is then compared to a model with equality constraints on the intercepts of only the items that are found metric invariant. Using MIs, one again identifies noninvariant intercepts sequentially until none of the MIs of the intercepts is larger than the cutoff, resulting in a partial scalar invariance model.[4]

When the number of items is large, specification search can involve substantial capitalization on chance (MacCallum et al., 1992). Consider an example by Skriner and Chu (2014), in which the authors evaluated measurement invariance of the 20-item Center for Epidemiologic Studies Depression Scale (CES-D) across four ethnic groups, before conducting latent mean comparisons. If we just consider invariance constraints on intercepts, there were 60 of them (i.e., number of items × [number of groups - 1]). Theoretically, there were 2^{60} possible partial invariance models the researcher could have ended up with, thus a lot of model uncertainty. For example, MacCallum et al. (1992) studied the problem of specification search in structural equation modeling, the underlying framework for the current discussion of factorial invariance. The authors showed that, for sample data sets simulated from the same data generating model, specification search may lead to the selection of very different models when the sample size is small to moderate (i.e., 400 or less). Therefore, there could be a lot of model uncertainty in a specification search to identify a partial invariance model. However, the standard errors (*SE*s) obtained in the final partial invariance model does not capture the model uncertainty due to the search and may underestimate the sampling variability of the parameter estimates, leading to inflated

—————

[4] Theoretically, the process is then repeated for identifying noninvariant unique factor variances. Practically, for continuous indicators, partial scalar invariance is sufficient for valid latent mean comparisons (Schmitt & Kuljanin, 2008; Whittaker, 2013), so in the current paper we stop the specification search process when the final partial scalar invariance model is reached.

Type I errors and suboptimal coverage rates of the sample confidence intervals (CIs).[5]

Previous research has shown that FS worked well in identifying biased items, although they usually relied on the design with two groups, six items, and high composite reliability of above .90 (Jung & Yoon, 2016; Yoon & Kim, 2014; Yoon & Millsap, 2007). Yoon and Kim (2014) showed, in a simulation study with six items and two groups, that sequential specification search performed well in terms of correctly identifying the biased items while keeping the false positive rate of flagging an invariant item to less than 3%. Jung and Yoon (2016) further suggested that using a more conservative cutoff of 6.635 on modification indices reduced the false detection rates. However, unlike conventional statistical inferences, when it comes to adjusting for partial invariance, false positives may be less costly than false negatives,[6] as demonstrated in Shi, Song, and Lewis (2017). On the other hand, Marsh et al. (2018) evaluated FS using the ΔCFI criterion (Cheung & Rensvold, 2002) in conditions of 15 groups with five items when *all* items were biased. They found that FS performed poorly in terms of biases and mean squared errors (MSEs) for the measurement and structural parameter estimates, compared to the alignment method to be discussed below. However, it should be noted that the change in CFI criterion resulted in more false negatives as it tends to keep constraints that may have large modification indices, especially in large samples. To our knowledge, however, there have been no previous studies evaluating parameter estimations following FS with sequentially-relaxed constraints based on MIs.

───────

[5] A similar problem has been widely studied in using stepwise regression for variable selections (Harrell, 2001), and an adjusted inference procedure that takes into account model selection uncertainty was proposed by Tibshirani et al. (2016).

[6] This is because incorrectly constraining unequal parameters biases the latent parameters, whereas allowing parameters that are indeed equal to be different only reduces the precision in the latent parameter estimates.

Approximate Invariance Methods

Alignment Optimization. Different from the specification search methods (e.g., FS) for noninvariant item detection, which involve many model comparisons when the number of items and number of groups are large, alignment optimization (AO) simplifies noninvariance detection by minimizing small noninvariances while retaining large noninvariances, and can be used across many groups (2–100; Asparouhov & Muthén, 2014; Muthén & Asparouhov, 2018). While AO still examines noninvariance under a factor model, it does not require a specific stage of factorial invariance (e.g., scalar invariance) for meaningful group comparisons; instead it aims to reach an *approximate invariance* condition that has the minimum number of highly noninvariant parameters across groups (Muthén & Asparouhov, 2018).

Specifically, under the one factor model in (2), the alignment algorithm first evaluates a configural model by standardizing latent factors across all K groups for identifications (i.e., $\alpha_k = 0$ and $\psi_k = 1$) and freely estimating factor loadings $\lambda_{k,0}$ and intercepts $\nu_{k,0}$; we denote this baseline configural model M_0. Given that there are infinitely many possible sets of $\{\alpha_k, \psi_k, \lambda_k, \nu_k\}$ that would give the same model-implied means and covariances for $\mathbf{y}$, for $k > 1$, AO obtains α_k and ψ_k that minimize a simplicity function (Asparouhov & Muthén, 2014)

$$F = \sum_j \sum_{k_1 < k_2} w_{k_1,k_2} f(\lambda_{jk_1,1} - \lambda_{jk_2,1}) + \sum_j \sum_{k_1 < k_2} w_{k_1,k_2} f(\nu_{jk_1,1} - \nu_{jk_2,1}), \qquad (3)$$

where $\lambda_{jk_1,1}$ and $\lambda_{jk_2,1}$ denote factor loadings of item j in groups k_1 and k_2 under model M_1; $\nu_{jk_1,1}$ and $\nu_{jk_2,1}$ denote the intercepts; $w_{k_1,k_2} = \sqrt{N_{k_1} N_{k_2}}$ is a weight coefficient based on the geometric mean of group sample sizes, and $f(\cdot)$ is some component loss function. Asparouhov and Muthén (2014) suggested a good choice of $f(\cdot)$ that prefers few large noninvariant parameters and many small pairwise differences in factor loadings and

intercepts (i.e., approximate invariance) is

$$f(x) = \sqrt{\sqrt{x^2 + \epsilon}}, \tag{4}$$

with a small ϵ such as .01 (which is the default in Mplus).

As a result, the AO model has the same fit as the baseline configural invariance model, but the parameter estimates, including the latent means and variances, are directly interpretable as they are set on a metric that is comparable across groups. If identification of noninvariant items is of interest, AO evaluates whether a measurement parameter in one group is different from those in other groups by iteratively identifying an invariant set for each parameter via pairwise comparison tests, with more details described in Asparouhov and Muthén (2014). Asparouhov and Muthén (2014) further distinguished between FIXED and FREE AO estimation. For FIXED AO, factor mean in the first group is $\alpha_1 = 0$; for FREE AO, it estimates α_1.

Given that AO is relatively new, there has been limited, but increasing, research efforts to evaluate its performance. Asparouhov and Muthén (2014) found that, for a model with five continuous items and at most one noninvariant item, AO yielded relatively unbiased parameter estimates across conditions of sample sizes (100 or 1,000) and numbers of groups (2 to 60), but with larger biases for conditions with small numbers of groups and sample size (e.g., 20% for latent mean estimates). They also compared AO with maximum likelihood and the Bayesian version of AO with noninformative priors (denoted as BAO in this paper), and found that coverage rates of 95% CI of AO and credible interval (CrI) of BAO were close to or above 95% except for some conditions with small number of groups. Muthén and Asparouhov (2014) suggested that, when there are many groups, AO can produce trustworthy results when the percentage of noninvariant parameters was less than 25%. Similarly, Flake and McCoach (2018), in a simulation study with a two-factor model and 14 polytomous items, found AO performed well in recovering measurement and

structural parameters with 29% or less noninvariance (bias $\leq$ 8%), but showed more biases (up to 23%) in conditions with 43% noninvariance. Marsh et al. (2018) found that AO outperformed specification search method using CFI in terms of bias and MSE of parameter estimates in a simulation with 15 groups and noninvariance on all items. Kim et al. (2017), in a simulation study with a one-factor model with six items, also supported the use of AO to correctly identifying noninvariant items with many groups (25 or 50). On the contrary, Pokropek et al. (2019) compared correctly specified partial invariance model, AO, and SV priors (without alignment) in conditions with large number of groups (24), large sample size (1,500), and small number of items (3 to 5), and found that AO only performed reasonably in recovering latent mean estimates with good coverage of 95% CIs when there were no more than 20% noninvariant items.

While previous results of AO were promising especially in many groups, it is not clear how it performs in a smaller number of groups with varying numbers of items and proportions of noninvariance. More importantly, as a majority of empirical studies evaluating measurement invariance focused on demographic subgroups, which involves only a few groups, the question remains how AO performs relative to specification search in recovering measurement and structural parameters and how this depends on other factors such as sample size.

Bayesian Approximate Invariance. In the context of factorial invariance, Muthén and Asparouhov (2013) proposed the use of small variance (SV) priors under the Bayesian framework, in which researchers assign $N(0, \sigma_d)$ priors for some small values of σ_d (e.g., $\sigma_d = 0.1$, corresponding to a variance of 0.01) on the *differences* of loadings and intercepts across groups. Such priors represent the a priori belief that approximate measurement invariance holds so that the differences in intercepts and loadings are relatively small, thus a less restrictive assumption than in partial invariance models where some loading and intercept differences are exactly zero (van de Schoot et al., 2013). In other words, approximate invariance allows a "wiggle room" in the form of small degree of

parameter differences across groups in an attempt to balance invariance assumption and model fit (van de Schoot et al., 2013). The use of SV priors is considered more realistic as exact scalar or strict invariance may be too ideal to reach (Marsh et al., 2018). However, as discussed in Pokropek et al. (2020), the choice of the scale value in SV priors could affect the coverage rates of the latent means, and the use of SV priors worked best when the prior scale approximately matched the actual variability of the noninvariant parameters in the data.

Muthén and Asparouhov (2013) performed a simulation study with six continuous items across 10 groups to evaluate the statistical power of using SV priors for detecting item biases. Based on the 95% credible interval (CrI) of the posterior distributions of loading/intercept differences, SV priors had high detection rates for small violations of noninvariance. They found, however, that the use of SV priors sometimes produced biased estimates of factor means and variances In another simulation study with a one-factor model and four continuous items for two groups, van de Schoot et al. (2013) found the use of SV priors with variances of 0.01 or 0.005 for intercept differences performed well in recovering latent mean parameters, but had overestimated standard errors due to the alignment issue as described in Muthén and Asparouhov (2013), which relates to the factor indeterminacy issue as the scale of the latent variable is not set (see also Levy & Mislevy, 2016). Both Asparouhov and Muthén (2014) and van de Schoot et al. (2013) suggested resolving this issue by using the alignment optimization technique, described in the previous section, together with SV priors. However, to our knowledge, this approach, which we denoted as BAIA (Bayesian approximate invariance with alignment) in this study, had not been systematically evaluated. Therefore, in the current paper, we also examine two methods based on Bayesian approximate invariance with SV priors and AO.

Bayesian Approximate Invariance with Alignment (BAIA).   To resolve the factor inderterminacy issue with the use of SV priors, both Asparouhov and Muthén (2014) and van de Schoot et al. (2013) suggested using the alignment optimization technique to

set the scale of the latent variable, an approach we denoted as BAIA in this study. The BAIA method is similar to BAO, but uses SV priors on the differences of the loadings and the intercepts across groups. As such, BAIA combines Bayesian approximate invariance with AO. Muthén and Asparouhov (2013) discussed several advantages of BAIA, including the possibility of modeling unique factor covariances with strong regularizing priors, better interpretability of alignment results, and stabilized alignment estimations. To our knowledge, however, BAIA has not been evaluated in the simulation studies by Muthén and Asparouhov (2013) nor in other previous studies, so our study will provide insights on whether combining SV priors and AO would be desirable.

Two-Step Bayesian Approximate Invariance with Alignment (BAIA-2S). Muthén and Asparouhov (2013) also proposed a two-step procedure, which runs Bayesian approximate invariance in the first step; however, rather than using the parameter estimates on the measurement and structural parameters from the approximate invariance model, the goal of the first step is to identify noninvariant loadings and intercepts using the pairwise comparisons tests described in Asparouhov and Muthén (2014). In the second step, a partial factorial invariance model is fit to the data with invariance constraints only on parameters that were not flagged as noninvariant in the first step, and like FS, the parameter estimates of this two-step approach will be from the partial invariance model in the second step.

In a simulation with 10 groups, Muthén and Asparouhov (2013) found that the two-step approach gave better factor means and variances estimates than BAO, although the CI coverage rates for the factor variances were suboptimal. The two-step approach by Muthén and Asparouhov (2013), however, did not incorporate the alignment method later developed. While to our knowledge this two-step approach with alignment—BAIA-2S—has not been evaluated in the literature, van de Schoot et al. (2013) suggested that including alignment helped solve the factor indeterminacy issue and further improves the estimates under Bayesian approximate invariance. Therefore, we include BAIA-2S in the list of

methods in the simulation.

Current Studies

In the current research, we compared the performance of forward specification search (FS) and four approximate invariance methods: frequentist alignment optimization (AO), Bayesian AO with noninformative priors (BAO), Bayesian approximate invariance with alignment (BAIA), and two-step BAIA (BAIA-2S), in recovering structural (i.e., α, ψ, and path coefficient) and measurement (i.e., λ and ν) parameters in small numbers of groups. Given inferences based on FS ignored the model uncertainty in the search, we expected it would yield *SE*s that are too small and CIs that are too narrow. On the other hand, by avoiding the instability in searching for a partial invariance model, the approximate invariance methods may yield parameter estimates with higher efficiency (as measured by MSE), as previously demonstrated in Marsh et al. (2018), for the latent parameters.

For each simulation study, we report how we determined our design conditions, the number of replications, and all evaluation measures of the simulation results.

Study 1

In Study 1, we compared five methods for adjusting item bias using a one-factor model with two groups, which is commonly seen in a variety of research. Given that previous simulation studies examining FS have only focused on relatively small numbers of items and two groups (Jung & Yoon, 2016; van de Schoot et al., 2013; Yoon & Kim, 2014), in Study 1 we expanded the design conditions to include larger numbers of items. The data generating model followed a factor model with latent means and variances of $\alpha_1 = 0$ and $\psi_1 = 1$ for Group 1 and $\psi_2 = 1.3$ for Group 2; α_2 was manipulated as one of the design conditions. All invariant loadings were chosen to be 0.7 and all invariant intercepts were 0. The loading and intercept patterns for noninvariant items were described in the design

factors below. In addition, to introduce model misspecification so that the simulated data resembled more closely to those in actual research, we used a procedure similar to MacCallum and Tucker (1991) by adding minor unique factor covariances with magnitudes between -0.1 and 0.1 to the data. The unique factor covariance matrix was simulated by first sampling from an LKJ distribution for correlation matrices (Lewandowski et al., 2009), and then dividing by 10 so that the maximum absolute value was 0.1. When fully invariant data were simulated, the analytic model had a population RMSEA of .057, which was around the typical value of an acceptable model in real research (Hu & Bentler, 1999).[7] The unique factor covariance matrix with $p = 6$ shown in the Appendix. For all conditions, the unique factor variance of each item was set as $1 - 0.1 - \lambda_{j1}^2$ for both groups so that the total variance $= 1$ for all items in Group 1.

Design Conditions

Sample Size Per Group (n).   The sample size was kept the same across groups with $n = 100$ or 500, which was similar to previous simulation studies (e.g., Yoon & Kim, 2014).

Number of Items (p).   The number of items was $p = 6$, 12, or 24. Yoon and Kim (2014) and Jung and Yoon (2016) showed that sequential specification search methods effectively identified biased items for models with $p = 6$, and most of the previous simulation studies had similarly small p (e.g., Asparouhov & Muthén, 2014; van de Schoot et al., 2013). With larger p the composite reliability increased, but there was also potentially more capitalization on chance as demonstrated in MacCallum et al. (1992).

---

[7] As pointed out by one reviewer, with misspecification in the sample model, the parameter estimates may converge to values (ϑ^*) that are different from the population values (ϑ). In our simulations, the impact of misspecification was relatively minor, as the differences between the ϑ^* and ϑ were no more than 0.021 for the latent mean and no more than 0.037 for the latent variance. Given that applied researchers are usually interested in the ϑ instead of ϑ^*, we used the former as reference when evaluating the sample estimates. Due to the unmodeled unique covariance, the loadings (but not the intercepts) also converged to slightly different values across groups (with a difference of about 0.01) even when they were invariant in the correctly specified model, so the LRT in FS may not follow its theoretical behavior (Yuan & Bentler, 2004). In our simulations, the impact was negligible as discussed in the follow-up analyses in Study 2.

Pattern of Noninvariant Parameters. We simulated four patterns of noninvariance based on the proportion of noninvariant *parameters* (r_{ni}; *not* the proportion of noninvariant items) and the direction of bias. Noninvariant loadings and intercepts were present simultaneously. Specifically, there were one condition with $r_{ni} = 0$, two conditions with $r_{ni} = 1/3$ (balanced vs. skewed directions), and one condition with $r_{ni} = 2/3$.[8] [9] When $r_{ni} = 1/3$, the first $p/3$ items (e.g., items 1 to 8 when $p = 24$) were simulated to have noninvariant loadings; the first $p/6$ items and items $p/2 + 1$ to $p/2 + p/6$ (e.g., items 1 to 4 and 13 to 16) were simulated to have noninvariant intercepts. With this pattern, $p/6$ items had both noninvariant loadings and intercepts, $p/2$ items had invariant loadings and intercepts, and the remaining items had either noninvariant loadings or noninvariant intercepts. When $r_{ni} = 2/3$, the first $2p/3$ items were simulated to have noninvariant loadings, and the first $p/3$ items and items $p/2 + 1$ to $p/2 + p/3$ were simulated to have noninvariant intercepts. As a result, only $p/6$ items had both invariant loadings and invariant intercepts.

In half of the conditions with $r_{ni} = 1/3$, the direction of bias was balanced across items; it was skewed in the other half. Specifically, for the 1/3-balanced noninvariance

---

[8] While it is not uncommon in practice to find a large proportion of noninvariant parameters, such as Hasan et al. (2019) who found that a 12-item modified Shortened Adapted Social Capital Assessment Tool had 2/3 noninvariant parameters (5 loadings, 11 intercepts) across gender, theoretically it is debatable whether the construct is still comparable when a majority of the items are found noninvariant. Whereas Steenkamp and Baumgartner (1998) and Vandenberg and Lance (2000) suggested that more than half of the items should be invariant for meaningful comparisons, the studies by Marsh et al. (2018), Pokropek et al. (2019), and Shi et al. (2019) assumed that latent variables were comparable with only a few invariant items or even with all items only approximately invariant. We follow Marsh et al. and Shi et al. to include simulation conditions with more than 50% noninvariant parameters, as in practice it is possible that many items may have slightly different intercepts and loadings; however, readers should interpret the simulation results with 2/3 noninvariant parameters with caution, and additional considerations such as magnitudes of noninvariance and conceptual meaning of the items should be used to decide whether the latent variable is still comparable across groups.

[9] As discussed in Asparouhov and Muthén (2014), for a hypothetical model in which most parameters are noninvariant in the same direction, there exists an equivalent but simpler model with fewer noninvariant parameters, which Asparouhov and Muthén referred to as the "alignment" issue. In this case, both FS and approximate invariance methods would identify noninvariant parameters according to the simpler model. Therefore, we only simulated data with many noninvariant parameters in balanced directions, which is similar to the design by Marsh et al. (2018).

condition, half of the noninvariant items had larger loadings (and intercepts) in Group 1 than in Group 2, whereas the other half had larger loadings (and intercepts) in Group 2. For the 1/3-skewed noninvariance condition, all noninvariant items had larger loadings (and intercepts) in Group 1 than in Group 2.

The magnitudes of noninvariance were chosen according to the effect size benchmarks from Nye et al. (2018) based on a review of the organizational literature. Specifically, half of the noninvariant parameters had a medium effect size (i.e., $\Delta\lambda = 0.2$, $\Delta\nu = 0.5$), whereas the other half had a small effect size (i.e., $\Delta\lambda = 0.1$, $\Delta\nu = 0.25$). Table 1 shows the loading and intercept pattern with $p = 12$ for the manipulated noninvariance patterns.

Latent Mean of Group 2 (α_2). We manipulated α_2 to be 0 and 0.5. The zero condition allowed evaluation of empirical Type I error rates, and 0.5 was a typical effect size value deemed minimally clinically important as suggested by Angst et al. (2017).

Analytic Models

We analyzed each simulated data set with five approaches as described below. Because the data were simulated with unique covariances, there were model misspecification in all five approaches as they assumed no unique covariances.

Forward Stepwise Specification Search (FS). As previously described, with FS, invariance constraints at a particular stage were sequentially freed based on modification indices, until none of the modification indices were above the 3.84 cutoff. We used R and the *lavaan* package (Version 0.6.7; Rosseel et al., 2020) to implement the FS procedure.

Alignment Optimization (AO). Maximum likelihood estimation was used with the AO method, using the `ALIGNMENT=FIXED` option in Mplus. We used the default $\epsilon = .01$ for the component loss function (see equation [4]).

Bayesian Alignment Optimization (BAO). BAO used MCMC (with Gibbs sampling) in Mplus (Version 8.3) with the `ESTIMATOR=BAYES` option. We used the default

noninformative priors in Mplus (i.e., uniform priors from $-\infty$ to ∞ or from 0 to ∞) for all parameters. The Gibbs sampling algorithm drew at least 10,000 samples, and stopped when the potential scale reduction factor dropped to below 1.01 (Vehtari et al., 2019). The first half of the posterior samples were discarded as burn-ins. As described in Asparouhov and Muthén (2014) and van de Schoot et al. (2013), alignment was then performed for each MCMC iteration (post burn-in) to improve interpretability using the `ALIGNMENT=FIXED` option.

Bayesian Approximate Invariance With Alignment (BAIA). Given that Nye et al. (2018) suggested that a difference in standardized loadings less than 0.1 and a difference in intercepts less than 0.25 were considered negligible, we specified a $N(0, 0.1)$ prior for the difference of each pairs of loadings and a $N(0, 0.25)$ prior for the difference of each pair of intercepts across groups.[10] Following Asparouhov and Muthén (2014), alignment was then performed using the `ALIGNMENT=FIXED(BSEM)` option.

Two-Step BAIA (BAIA-2S). From the output of BAIA, we obtained the Mplus test results that indicated which loadings and intercepts were found noninvariant, based on pairwise Wald tests with significance level at .001 (as described in Asparouhov & Muthén, 2014). We then used *lavaan* to specify a partial invariance model according to the Mplus output.

In addition, we fitted the partial scalar invariance model with the correct constraints on the loadings and intercepts (but not for the unique factor covariances), denoted as PI.

Data Generation

We used R and *lavaan* to simulate 2,500 data sets for each condition, which would be sufficient to keep the Monte Carlo error to less than 2% of the *SE*s of the parameters. For Type I errors, 2,500 replications corresponded to a margin of error of less than 0.5

---

[10] This corresponds to `N(0, 0.01)` and `N(0, 0.0625)` in Mplus, as there the second number requires input of the variance instead of the standard deviation.

percentage points, which satisfied the stringent criterion defined by Bradley (1978). The `simulateData()` function in *lavaan* was used to simulate normally distributed data for two groups.

Each simulated data set was analyzed using the methods described above. For models fitted in Mplus, the results were then imported into R using the *MplusAutomation* package (Version 0.8; Hallquist & Wiley, 2018). For all methods, the models were identified by fixing $\alpha_1 = 0$ and $\psi_1 = 1$. In each analysis we obtained point (maximum likelihood or posterior median), uncertainty (standard error or posterior *SD*), and interval (95% Wald CI or symmetric Bayesian CrI) estimates for α_2 and ψ_2, the latent mean and variance of Group 2, as well as for all loadings (λs) and intercepts (νs) of both groups.

Evaluation Criteria

For each parameter ϑ and analytic method in each simulation condition, we evaluated the following.

Bias. The bias was computed as

$$\bar{\hat{\vartheta}} - \vartheta,$$

where $\bar{\hat{\vartheta}} = \frac{\sum_{r=1}^{R} \hat{\vartheta}_r}{R}$ is the mean of the $\hat{\vartheta}_r$ estimates across 2,500 replications, and ϑ is the population parameter value. We only computed bias for α_2 and ψ_2, as biases for the λs and νs might cancel out, and estimation performance for these parameters was better captured by the relative efficiency.

Relative Efficiency. An unbiased estimator may not be preferred over a slightly biased estimator if the former has a much larger sampling variance. This is particularly true for Bayesian methods, as Bayes estimators tend to be slightly biased but may have smaller sampling variability, so that overall it tends to be closer to the true value on average. To compare the different methods balancing the bias-variance tradeoff, we computed the

relative efficiency (RE) of each method relative to the estimator based on the correctly specified partial invariance model (PI). Specifically, for an estimator based on method M,

$$\mathrm{RE}(\hat{\vartheta}^M, \hat{\vartheta}^{\mathrm{PI}}) = \frac{\mathrm{MSE}(\hat{\vartheta}^{\mathrm{PI}})}{\mathrm{MSE}(\hat{\vartheta}^M)} \tag{5}$$

$$= \frac{\sum_{r=1}^{R}(\hat{\vartheta}_r^{\mathrm{PI}} - \bar{\hat{\vartheta}}^{\mathrm{PI}})^2/R}{\sum_{r=1}^{R}(\hat{\vartheta}_r^M - \bar{\hat{\vartheta}}^M)^2/R}, \tag{6}$$

where MSE is the mean squared error. An RE larger than one means that method M should be preferred over PI. For λs and νs, we computed REs as ratios of average MSEs across p items.

Error Rate of 95% CI and CrI.   For each method and simulated data set we obtained either (for maximum likelihood) the 95% Wald CI by $\hat{\vartheta} \pm z_{.975}\hat{SE}(\vartheta)$, where $z_{.975}$ is the 97.5th percentile in a standard normal distribution, or (for Bayesian) the 95% symmetric CrI as the 2.5th and the 97.5th percentiles of the posterior distribution. The empirical error rates, denoted as α^*, was calculated as the proportion of times the constructed CI or CrI failed to contain the population ϑ value (i.e., $1 -$ coverage rates). A valid 95% CI or CrI should have an error rate of at most 5%.

We used the *SimDesign* (Version 2.2; Chalmers & Adkins, 2020) package in R to structure the simulation studies.

Results

The simulation results are shown in Figure 1 (bias), Figure 2 (RE), and Figure 3 (CI/CrI error rate). Here we summarize the results for each approach for adjusting partial invariance.

FS.   The performance of FS, in terms of absolute bias, RE, and error rate, was sensitive to the amount of noninvariant parameters (r_{ni}) and sample size. Biases of FS were close to zero for most conditions (median bias $= 0.04$). However, for αs, FS yielded the

highest biases (up to 0.29) among tested methods for the condition with 2/3-balanced noninvariance pattern. Overall, FS maintained good REs for νs, α, and ψ for most conditions, except with $r_{ni} = 2/3$. For λs, FS showed lower REs (median RE = 0.73), but they were still comparable to AO and BAO. When $r_{ni} = 2/3$ and $n = 500$, FS demonstrated inflated error rates for all parameters (median = 37.15%) and decreased REs. The increase in bias and error rates was most prominent in estimating α (with error rates up to 83.48%); the decrease in efficiency was most noticeable in estimating νs and αs.

AO. Compared to all other tested methods, AO in general gave the most desirable performance in terms of bias, RE, and error rate. AO well controlled the bias close to zero for most conditions (median bias = 0.03). When the pattern of noninvariance was 1/3-skewed, although biases of AO were substantially above zero for ψs, other methods yielded comparable or even larger biases. AO maintained good REs (median RE = 0.90) and error rates close to the 5% nominal level (median error rate = 5.88%) across varying degrees of noninvariance and sample sizes. For conditions with the 1/3-skewed noninvariance pattern, AO gave slightly worse estimates of α and ψ, especially with larger sample size ($n = 500$).

BAO. The performance of BAO was similar to AO, in terms of error rate and bias, but was more sensitive to sample size. BAO, just as AO, controlled bias close to zero for most conditions (median bias = 0.04), except for ψs when the noninvariance pattern was 1/3-skewed. When estimating λs, it consistently had lower REs particularly when $n = 100$, but had higher REs than AO and generally outperformed the correctly specified PI (i.e., RE > 1) in small samples for α and ψ. The error rates for CrIs were higher for λ with $p = 24$ and $n = 100$ (median error rate = 17.45%), and for α with the 1/3-skewed noninvariance pattern (median error rate = 14.44%), than for other conditions.

BAIA. Similar to BAO, BAIA was consistently less efficient in yielding estimates of λs, and was even less efficient when sample size was small ($n = 100$), but had similar REs as BAO when $n = 100$ for α, and better REs for those parameters when $p = 6$. In terms of

bias, it produced unstable estimates of ψs across conditions. In terms of error rates, it performed poorly in small samples for λs and ψ, and in large samples for α.

BAIA-2S. The performance of BAIA-2S had a similar pattern as FS in terms of bias and RE. BAIA-2S had biases close to zero for most cases but increased biases for the conditions with 1/3-skewed and 2/3-balanced noninvariance pattern. It also had good REs when the data were fully invariant but gradually dropped off when r_{ni} increased. Like FS, it generally had better performance for estimating νs and α in larger samples, but better performance for estimating λ and ψ in small samples. The inflation of error rates of CIs increased with larger r_{ni}, but to a lesser degree than FS.

Overall, AO showed the best performance in terms of bias and RE across parameters and in terms of maintaining good coverage rates of CIs, but BAO yielded estimates with higher precision in small samples. FS also performed well for most conditions as long as the proportion of noninvariant parameters was no more than 1/3, but it had a substantial drop-off with a larger proportion of noninvariant parameters. Because the data were simulated with misspecification in the covariance structure, the results also showed that FS was more sensitive to misspecification with increased error rates in λ and ψ, whereas AO and BAO were more robust.

We also did a follow-up analysis to compare parameter bias between invariant and noninvariant loadings and intercepts for the four conditions with $N = 500$, $p = 12$, and 1/3-balanced and 2/3-balanced noninvariance patterns. Whereas the absolute parameter bias was generally larger for noninvariant than invariant loadings and intercepts, in general it was smallest for AO (average absolute bias = 0.01 and 0.02 for invariant and noninvariant loadings) than for FS (average absolute bias = 0.02 and 0.05 for invariant and noninvariant loadings). The biases of FS were similar to those of BAIA-2S, and were larger than those of BAO and BAIA. The same pattern was observed for both $r_{ni} = 1/3$ and $r_{ni} = 2/3$. Full results can be found in the supplemental material.

Study 2

In Study 2, we attempted to replicate the findings in Study 1 in a four-group setting. The design factors were the same as in Study 1, except that we excluded the $p = 24$ conditions, and used $n = 50$ or 500 per group as the sample size conditions. The two conditions of α patterns were $\{0, 0, 0, 0\}$ and $\{0, .5, -.25, .125\}$. For all conditions, ψs had the pattern $\{1, 1.3, 1, 1.3\}$. In addition, whereas in Study 1 we simulated some items with small bias and some with medium bias, in Study 2 we simulated those items with either small bias or *large* bias (i.e., 0.3 in loadings and 0.75 in intercepts). Because there were three freely estimated αs and ψs, we calculated REs using the average MSEs and the average error rates of CIs (or CrIs), the same way we did for λs and νs in Study 1.

As shown in Figures 4 and 5, the results were highly similar to those in Study 1, in that AO overall gave good REs across conditions and with good CI error rates. However, with four groups BAO performed almost identically to AO in large samples but outperformed AO in small samples for νs, αs, and ψs, with similar error rates. Similar to Study 1, FS and BAIA-2S performed similarly except when there were 2/3 noninvariant items, where BAIA-2S was better. The small variance priors in BAIA resulted in increased REs for estimating αs when $n = 50$, $p = 12$, and all αs $= 0$, but generally showed poor REs for estimating λs and ψs when $n = 50$ and $p = 12$ with highly inflated error rates. Overall, with four groups, BAO and AO should be the methods of choice for adjusting partial invariance, with BAO being more efficient in small samples.

For the 16 conditions with equal αs across groups (i.e., $\alpha_k = 0$ for all k), we also obtained the empirical Type I error rates for the omnibus test $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$. The likelihood ratio test (LRT) with three degrees of freedom was used to compare the model without constraints on the αs and that with equality constraints on the αs for the correctly specified PI, FS, and BAIA-2S, whereas the Wald test was used for AO (as LRT was not available with alignment). Significance tests were not available from BAO and BAIA in

Mplus at the time of writing.

As expected and shown in Table 2, when $r_{ni} = 0$, all methods yielded reasonable empirical Type I error rates (α^*s), but AO was slightly conservative in small samples (1.4% to 2.0%). With $r_{ni} = 1/3$, AO generally maintained acceptable α^* except with the 1/3-skewed noninvariance pattern and $n = 500$ (9.0%), whereas both FS and BAIA-2S had inflated α^*s between 6.4% and 10.6%. With $r_{ni} = 2/3$, both FS and BAIA-2S had highly inflated α^* ($> 75\%$ for FS, $> 17\%$ for BAIA-2S), while AO performed much better but still had inflated α^* of 8% to 10% when $n = 500$.

When Score Reliability Is Low.   As the simulation conditions in Study 2 had high composite reliability with Cronbach's alphas $= .852$ when $p = 6$ and $.920$ when $p = 12$, we would like to investigate whether the results would change when the composite reliability is low. Thus, we simulated additional data as in Study 2 but with the standardized loadings reduced to .4, resulting in Cronbach's alphas $= .533$ when $p = 6$ and $.696$ when $p = 12$. The pattern of the results was in general consistent with conditions with high reliability, so we only reported a few key findings (with the full results available in the supplemental material). Specifically, while FS had worse REs when there were 2/3 noninvariant parameters, it performed as good as AO in other conditions, and had better REs in estimating λs and ψs. However, FS suffered from even higher error rates of CIs in low reliability conditions, so while they produced good parameter estimates, the standard errors were underestimated as it did not take into account the model uncertainty in the search process. Across conditions with low reliability, BAO had consistently higher REs than AO, especially $p = 6$ (i.e., when reliability is low), and maintained good error rates except for some conditions of $n = 500$ and $p = 12$ (with median error rate $= 7.2\%$), while AO had more acceptable error rates across conditions. Both BAIA and BAIA-2S also had unacceptably high error rates (up to 99.8% and 40.5% in some conditions). Therefore, BAO should be preferred in small sample size and low reliability situations, while both AO and BAO should be preferred over FS if researchers want to obtain valid inferences.

When there were no misspecification.   As one of the reviewers pointed out, our finding that AO performed better than FS may be sensitive to model misspecification. To examine that, we run an additional simulation using the same conditions as in Study 2, except that there were no unique covariances in the data generating model. The results were essentially identical to those of Study 2, and readers can find more details in the supplemental material.

Study 3

In Study 3, we compare the specification search and approximate invariance approaches to adjust for partial invariance in estimating path coefficients and their differences across groups (i.e., interaction effects). As noted in Marsh et al. (2018), currently both AO and BAO can only be used for confirmatory factor analytic (CFA) models, meaning that one cannot simultaneously do alignment and estimate structural relations of latent variables with other predictors or outcomes. However, because the alignment model is just one of the infinitely many possible configural invariance model, Marsh et al. (2018) proposed the alignment-within-CFA (AwC) approach by including covariates or outcome variables in the second step and constrained the loading and intercept of one anchor variable to be the same as the alignment solution for each latent factor. By doing so, the definition of the latent variable would generally be highly similar to the alignment solution. Marsh et al. (2018) showed that the alignment method with maximum likelihood performed well in recovering latent means for 15 groups; however, to our knowledge no previous studies had directly investigated the performance of AwC in recovering structural coefficients.

Specifically, in Study 3, we evaluated various approaches to estimating path coefficients of a latent predictor on an observed outcome for three groups. We denoted the path coefficient as β_1. We compared the specification search approach that identifies a partial invariance model in the first step and includes the outcome as the second step, as

well as several variants of the AwC methods, including AO, BAO, and BAIA. In addition, we also included two approaches discussed in Muthén and Asparouhov (2013), including (a) BAIA-2S as in Studies 1 and 2, in which a partial invariance model was identified using BAIA in step 1, and (b) a one-step simultaneous Bayesian method with small variance priors on the loading and intercept differences and with the outcome variable also included in the model, but without alignment optimization, which we denoted as BAI.

We generated data similar to those in Studies 1 and 2 but with an additional outcome variable. Specifically, we simulated data with three groups in three scenarios: $\boldsymbol{\beta}_1 = \{0, 0, 0\}$, $\boldsymbol{\beta}_1 = \{.3, .3, .3\}$, and $\boldsymbol{\beta}_1 = \{.3, .1, -.1\}$. The simulation conditions in Study 3 were similar to those of Studies 1 and 2, with a 3 ($\boldsymbol{\beta}_1$) × 2 ($N = 50$ or 500) × 4 (fully invariant, 1/3-balanced, 1/3-skewed, or 2/3 noninvariant parameters) design. Here we only studied $p = 12$ as Study 1 showed that the number of items had little impact on the results. Similar to Study 1, we introduced minor unique factor covariances among the 12 items as well as the external outcome variable by randomly simulating a covariance matrix from an LKJ distribution divided by 10. For each condition and method, we evaluated the relative efficiency (RE) and error rate of the point and interval estimates. In addition, for the 16 conditions with equal β_1 values across groups (i.e., with no interactions), we evaluated the empirical Type I error rates.

Results

Given that the results for the loadings, intercepts, latent means, and latent variances were highly similar to those in Studies 1 and 2, we only presented the results for the latent regression coefficient estimates. As shown in Figure 6, in general, AO performed the best with REs close to 1 for all conditions with $r_{\text{ni}} = 0$ and with the 1/3-balanced noninvariance pattern, and REs > .90 for conditions with 1/3-skewed and 2/3 noinvariance patterns. FS performed similarly well with only a slight drop-off with 1/3-skewed and 2/3 noinvariance patterns. BAIA-2S generally also performed well with REs above 0.86. BAO had good

performance only when $n = 500$, whereas both BAI (without alignment) and BAIA (with alignment) both performed poorly in most conditions.

As shown in Figure 7, in terms of CI (or CrI) error rate, all methods generally had error rates less than 6% except for BAI, with AO (median = 5.27%) performing the best overall and FS performing reasonably well except in conditions with $r_{ni} = 2/3$. Overall, FS showed better performance for recovering latent regression coefficient parameters than for latent means.

Table 3 shows the Type I error rates (α^*) for detecting differences in β_1 across groups. All methods yielded values close to 5% and below the 7.5% benchmark (Bradley, 1978) for all conditions, except for one condition for BAIA and for BAIA-2S when $r_{ni} = 2/3$ (7.6% and 8.3%, respectively). Overall, BAO had the best control on α^* (3.8% to 6.1%) across conditions.

Discussion

Traditionally, researchers have relied on comparing multiple models to evaluate factorial invariance, and use sequential specification search methods to relax invariant constraints to obtain adjusted estimations and inferences on measurement and latent structural parameters. While previous research has generally supported the use of specification search with two groups, a small number of items, and high reliability conditions (Jung & Yoon, 2016; Yoon & Kim, 2014), the corresponding results ignored the model uncertainty induced in the specification search process as researchers might identify the wrong set of items. On the other hand, newer approaches, such as the Bayesian approximate invariance and the alignment optimization methods, showed promising results for large number of groups (Asparouhov & Muthén, 2014; Kim et al., 2017; Marsh et al., 2018), but it was unclear whether they had advantages over the specification search method in smaller number of groups. To provide more concrete recommendations for

applied researchers interested in valid comparisons of latent variables across a few groups (i.e., 2 to 4 groups), in this paper we conducted three simulation studies. To our knowledge the current study was first in comparing forward specification search (FS), alignment optimization with both maximum likelihood (AO) and Bayesian estimation (BAO), and Bayesian approximate invariance methods (BAIA and BAIA-2S), in terms of the estimation and inferences of model parameters. In addition, we also examined the alignment-within-CFA approaches proposed by Marsh et al. (2018) in comparison to FS for estimating latent regression coefficients while adjusting for partial invariance.

Summary of Simulation Results

Across Studies 1 to 3, we found, consistent with Yoon and Kim (2014), that FS performed well when the proportion of noninvariant parameters was no more than 1/3, but its performance rapidly dropped off in terms of relative efficiency and CI error rate with 2/3 noninvariant parameters, especially for measurement intercepts and latent means. Whereas previous studies showed that the use of a correctly specified partial invariance model could produce accurate parameter estimates with only a small proportion of invariant items (e.g., Pokropek et al., 2019; Shi, Song, & Lewis, 2017), when model uncertainty is taken into account, which represents a more realistic situation in practice, FS only works when the proportion of noninvariant parameters is relatively small ($\leq 1/3$) and when the reliability of the item scores is relatively high (i.e., $> .80$).

On the other hand, our studies show that both AO and BAO are at least as good as FS in terms of RE and error rate for most conditions and are much better with 2/3 noninvariant parameters. In small samples ($n = 100$ in Study 1 and $n = 50$ in Study 2), the use of Bayesian methods also makes BAO more efficient in estimating latent means and variances. Thus, whereas previous literature has shown that AO works well in large number of groups with relatively few large noninvariant parameters (e.g., Flake & McCoach, 2018; Kim et al., 2017; Marsh et al., 2018; Pokropek et al., 2019), the current

results show AO and BAO also work well in two to four groups.

The use of small variance priors with alignment, on the contrary, generally biases the parameter estimates, which is consistent with Muthén and Asparouhov (2013) and Pokropek et al. (2019), and leads to inflated error rates in credible intervals. Whereas in some conditions it provides more efficient latent mean estimates in small sample conditions, its performance is inconsistent across conditions.

Recommendations for Research

Based on the simulation results, if researchers' goal is to obtain valid and efficient comparisons on the means or regression coefficients across groups, we recommend the use of AO for both measurement and structural parameters. In the case of small samples (i.e., $n \leq 100$ per group), we also recommend the use of BAO with noninformative priors as it gives more efficient parameter estimates of latent means and variances (but not latent regression coefficients). The traditional FS approach generally also works well for estimating intercepts, latent means, and latent regression coefficients, as long as researchers are confident that there are no more than 1/3 noninvariant parameters and the reliability is high, but it should be cautioned that FS may be more sensitive to model misspecifications than AO and BAO.

On the other hand, our results do not inform the choice between FS and AO when researchers are interested in detecting the source of noninvariance. For research problems involving many groups, readers should consult Kim et al. (2017); future research is needed to compare the performance of FS and AO for detecting noninvariant items in few groups.

Limitations

As with other studies, there are several limitations in the current study that may limit the generalizability of the results and require further investigations in future studies.

First, as with the majority of previous simulation studies, our data generating models only had one latent variable. Whereas Flake and McCoach (2018) had shown that the alignment optimization method (AO) performed well with two correlated factors, future research is needed to evaluate how specification search and Bayesian approximate invariance methods performed with multiple latent factors in adjusting latent means and latent regression coefficients among factors. Second, we only simulated continuous indicators that satisfied the normality assumption, and it is unclear to what degree results may be similar or different with nonnormal and categorical indicators. Yoon and Kim (2014) suggested that specification search had somewhat higher false positive rates with polytomous data, while Flake and McCoach (2018) found AO performed well with polytomous data in large-scale studies. However, more comprehensive comparisons that include Bayesian methods is needed for nonnormal data.

Third, in our simulation studies, we assumed that researchers did not have any substantive knowledge in terms of which items were biased. While this is generally the norm in the existing measurement invariance literature (e.g., Schmitt & Kuljanin, 2008), both specification search and Bayesian methods with small variance priors can incorporate researcher knowledge. In the case of specification search, researchers can choose to only evaluate a subset of the items for noninvariance; in the case of small variance priors, researchers can place such priors only on a subset of items they consider close to invariant. The performance of these two methods may improve with the incorporation of such knowledge, and future studies can further evaluate this possibility.

Finally, while Muthén and Asparouhov (2014) previously suggested a rule of thumb that AO worked well in recovering the ordering of the means with 25% or less noninvariant parameters when the number of groups is large (e.g., > 20), our simulation results showed that in situations with few groups, AO works reasonably well with 33% noninvariant parameters, or with 67% noninvariant parameters provided that the direction of noninvariance balanced out approximately. Our results are thus more consistent with the

findings from Marsh et al. (2018), who found that AO worked well with 15 groups, five items, and 100% noninvariant items (most of which were small noninvariances, but some were large). This suggested that the performance of AO may depend on both the number of groups and the magnitudes of noninvariance, and future studies are needed to further refine the rule of thumbs of when AO may be reasonable to use.

Open Practices Statement

The R and Mplus codes for all simulation studies are available on the Open Science Framework (https://osf.io/y98sj/).

References

Angst, F., Aeschlimann, A., & Angst, J. (2017). The minimal clinically important difference raised the significance of outcome effects above the statistical level, with methodological implications for future studies. *Journal of Clinical Epidemiology, 82*, 128–136. https://doi.org/10.1016/j.jclinepi.2016.11.016

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(4), 495–508. https://doi.org/10.1080/10705511.2014.919210

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*, 144–152. https://doi.org/10.1111/j.2044-8317.1978.tb00581.x

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., & Adkins, M. C. (2020). Writing effective and reliable Monte Carlo simulations with the SimDesign package. *The Quantitative Methods for Psychology, 16*(4), 248–280. https://doi.org/10.20982/tqmp.16.4.p248

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Cheung, G. W., & Lau, R. S. (2012). A direct comparison approach for testing measurement invariance. *Organizational Research Methods, 15*(2), 167–198. https://doi.org/10.1177/1094428111421987

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Flake, J. K., & McCoach, D. B. (2018). An investigation of the alignment method with polytomous indicators under conditions of partial measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(1), 56–70. https://doi.org/10.1080/10705511.2017.1374187

Guenole, N., & Brown, A. (2014). The consequences of ignoring measurement invariance for path coefficients in structural equation models. *Frontiers in Psychology, 5.* https://doi.org/10.3389/fpsyg.2014.00980

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling*, 1–18. https://doi.org/10.1080/10705511.2017.1402334

Harrell, F. E., Jr. (2001). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis.* Springer.

Hasan, M. Z., Leoutsakos, J.-M., Story, W. T., Dean, L. T., Rao, K. D., & Gupta, S. (2019). Exploration of factor structure and measurement invariance by gender for a modified Shortened Adapted Social Capital Assessment Tool in India. *Frontiers in Psychology, 10*, 2641. https://doi.org/10.3389/fpsyg.2019.02641

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

Hsiao, Y.-Y., & Lai, M. H. C. (2018). The impact of partial measurement invariance on testing moderation for single and multi-level data. *Frontiers in Psychology, 9.* https://doi.org/10.3389/fpsyg.2018.00740

Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Jung, E., & Yoon, M. (2016). Comparisons of three empirical methods for partial factorial invariance: Forward, backward, and factor-ratio tests. *Structural Equation Modeling:*

A Multidisciplinary Journal, 23(4), 567–584.

https://doi.org/10.1080/10705511.2015.1138092

Kim, E. S., Cao, C., Wang, Y., & Nguyen, D. T. (2017). Measurement invariance testing with many groups: A comparison of five approaches. *Structural Equation Modeling: A Multidisciplinary Journal, 24*(4), 524–544.

https://doi.org/10.1080/10705511.2017.1304822

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford.

Levy, R., & Mislevy, R. J. (2016). *Bayesian psychometric modeling.* CRC Press.

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis, 100*(9), 1989–2001. https://doi.org/10.1016/j.jmva.2009.04.008

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111*(3), 490–504. https://doi.org/10.1037/0033-2909.111.3.490

MacCallum, R. C., & Tucker, L. R. (1991). Representing sources of error in the common-factor model: Implications for theory and practice. *Psychological Bulletin, 109*(3), 502–511. https://doi.org/10.1037/0033-2909.109.3.502

Marsh, H. W., Guo, J., Parker, P. D., Nagengast, B., Asparouhov, T., Muthén, B., & Dicke, T. (2018). What to do when scalar invariance fails: The extended alignment method for multi-group factor analysis comparison of latent means across many groups. *Psychological Methods, 23*(3), 524–545. https://doi.org/10.1037/met0000113

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13*(2), 127–143.

https://doi.org/10.1016/0883-0355(89)90002-5

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525–543. https://doi.org/10.1007/BF02294825

Muthén, B., & Asparouhov, T. (2013). BSEM measurement invariance analysis. http://www.statmodel.com/examples/webnotes/webnote17.pdf

Muthén, B., & Asparouhov, T. (2014). IRT studies of many groups: The alignment method. *Frontiers in Psychology*, *5*(AUG), 1–7. https://doi.org/10.3389/fpsyg.2014.00978

Muthén, B., & Asparouhov, T. (2018). Recent methods for the study of measurement invariance with many groups: Alignment and random effects. *Sociological Methods & Research*, *47*(4), 637–664. https://doi.org/10.1177/0049124117701488

Nye, C. D., Bradburn, J., Olenick, J., Bialko, C., & Drasgow, F. (2018). How big are my effects? examining the magnitude of effect sizes in studies of measurement equivalence. *Organizational Research Methods*, *22*(3), 678–709. https://doi.org/10.1177/1094428118761122

Pokropek, A., Davidov, E., & Schmidt, P. (2019). A Monte Carlo simulation study to assess the appropriateness of traditional and newer approaches to test for measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(5), 724–744. https://doi.org/10.1080/10705511.2018.1561293

Pokropek, A., Schmidt, P., & Davidov, E. (2020). Choosing priors in Bayesian measurement invariance modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–15. https://doi.org/10.1080/10705511.2019.1703708

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

Rosseel, Y., Jorgensen, T. D., & Rockwood, N. (2020). *Lavaan: Latent variable analysis* [R package version 0.6-7]. http://lavaan.org

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review*, *18*(4), 210–222. https://doi.org/10.1016/j.hrmr.2008.03.003

Shi, D., Song, H., DiStefano, C., Maydeu-Olivares, A., McDaniel, H. L., & Jiang, Z. (2019). Evaluating factorial invariance: An interval estimation approach using bayesian structural equation modeling. *Multivariate Behavioral Research*, *54*(2), 224–245. https://doi.org/10.1080/00273171.2018.1514484

Shi, D., Song, H., & Lewis, M. D. (2017). The impact of partial factorial invariance on cross-group comparisons. *Assessment*, *26*(7), 1217–1233. https://doi.org/10.1177/1073191117711020

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*(4), 430–444. https://doi.org/10.1080/00273171.2017.1306432

Skriner, L. C., & Chu, B. C. (2014). Cross-ethnic measurement invariance of the SCARED and CES-d in a youth sample. *Psychological Assessment*, *26*(1), 332–337. https://doi.org/10.1037/a0035092

Sörbom, D. (1989). Model modification. *Psychometrika*, *54*(3), 371–384. https://doi.org/10.1007/BF02294623

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research [Publisher: Oxford Academic]. *Journal of Consumer Research*, *25*(1), 78–90. https://doi.org/10.1086/209528

Thurstone, L. L. (1947). *Multiple factor analysis.* University of Chicago.

Tibshirani, R. J., Taylor, J., Lockhart, R., & Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, *111*(514), 600–620. https://doi.org/10.1080/01621459.2015.1108848

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthén, B. (2013). Facing off with scylla and charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00770

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70. https://doi.org/10.1177/109442810031002

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2019). Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of mcmc.

Whittaker, T. A. (2013). The impact of noninvariant intercepts in latent means models. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(1), 108–130. https://doi.org/10.1080/10705511.2013.742397

Yoon, M., & Kim, E. S. (2014). A comparison of sequential and nonsequential specification searches in testing factorial invariance. *Behavior Research Methods*, *46*(4), 1199–1206. https://doi.org/10.3758/s13428-013-0430-2

Yoon, M., & Millsap, R. E. (2007). Detecting violations of factorial invariance using data-based specification searches: A monte carlo study. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 435–463. https://doi.org/10.1080/10705510701301677

Yuan, K.-H., & Bentler, P. M. (2004). On chi-qquare difference and z tests in mean and covariance structure analysis when the base model is misspecified. *Educational and Psychological Measurement*, *64*(5), 737–757. https://doi.org/10.1177/0013164404264853

Table 1

Factor Loadings and Intercepts for Different Noninvariance Patterns in the Data Generating Model of Study 1.

Parameter	$r_{\text{ni}} = 0$ G1 & G2	$r_{\text{ni}} = 1/3$, balanced G1	G2	$r_{\text{ni}} = 1/3$, skewed G1	G2	$r_{\text{ni}} = 2/3$ G1	G2
λ_1	0.700	0.800	0.600	0.800	0.600	0.800	0.600
λ_2	0.700	0.650	0.750	0.750	0.650	0.650	0.750
λ_3	0.700	0.800	0.600	0.800	0.600	0.800	0.600
λ_4	0.700	0.650	0.750	0.750	0.650	0.650	0.750
λ_5	0.700	0.700	0.700	0.700	0.700	0.800	0.600
λ_6	0.700	0.700	0.700	0.700	0.700	0.650	0.750
λ_7	0.700	0.700	0.700	0.700	0.700	0.800	0.600
λ_8	0.700	0.700	0.700	0.700	0.700	0.650	0.750
λ_9	0.700	0.700	0.700	0.700	0.700	0.700	0.700
λ_{10}	0.700	0.700	0.700	0.700	0.700	0.700	0.700
λ_{11}	0.700	0.700	0.700	0.700	0.700	0.700	0.700
λ_{12}	0.700	0.700	0.700	0.700	0.700	0.700	0.700
ν_1	0.000	-0.250	0.250	0.250	-0.250	-0.250	0.250
ν_2	0.000	0.125	-0.125	0.125	-0.125	0.125	-0.125
ν_3	0.000	0.000	0.000	0.000	0.000	-0.250	0.250
ν_4	0.000	0.000	0.000	0.000	0.000	0.125	-0.125
ν_5	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ν_6	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ν_7	0.000	-0.250	0.250	0.250	-0.250	-0.250	0.250
ν_8	0.000	0.125	-0.125	0.125	-0.125	0.125	-0.125
ν_9	0.000	0.000	0.000	0.000	0.000	-0.250	0.250
ν_{10}	0.000	0.000	0.000	0.000	0.000	0.125	-0.125
ν_{11}	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ν_{12}	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Note. λ = factor loading. ν = measurement intercept. r_{ni} = proportion of noninvariant parameters.

Table 2

Percentage Empirical Type I Error Rates for Latent Mean Differences in Study 2.

Noninvariance Pattern	N	p	PI	FS	AO	BAIA-2S
0	50	6	5.7	5.8	1.4	5.7
		12	4.9	5.0	2.0	4.9
	500	6	5.5	5.8	5.1	5.6
		12	4.8	4.8	4.4	4.8
1/3-balanced	50	6	6.6	**9.5**	2.1	**9.3**
		12	5.4	7.2	2.8	6.6
	500	6	5.1	**8.0**	5.2	7.4
		12	4.5	7.0	5.5	6.4
1/3-skewed	50	6	5.2	**10.6**	1.7	**10.4**
		12	5.7	**8.1**	3.2	**8.8**
	500	6	5.1	**9.2**	6.7	7.2
		12	5.0	7.4	**9.0**	**7.6**
2/3	50	6	5.7	**19.9**	1.5	**13.9**
		12	4.8	**15.0**	2.7	**9.0**
	500	6	5.5	**75.6**	**7.8**	**17.4**
		12	5.1	**75.2**	**10.4**	**18.2**

Note. p = number of items. PI = correctly specified partial scalar invariance model. FS = forward stepwise specification search. AO = alignment optimization. BAIA-2S = two-step Bayesian approximate invariance with alignment. Type I error rates larger than 7.5% are bolded.

Table 3

Percentage Empirical Type I Error Rates for Latent Mean Differences in Study 3.

Noninvariance Pattern	β_1	N	PI	FS	AO	BAO	BAIA	BAIA-2S
0	{0, 0, 0}	100	6.0	6.0	6.2	6.1	6.2	6.0
		500	4.5	4.5	4.5	4.5	4.5	4.5
	{0.3, 0.3, 0.3}	100	6.1	6.1	4.7	4.8	4.9	6.1
		500	4.4	4.4	3.8	3.8	3.9	4.4
1/3-balanced	{0, 0, 0}	100	5.0	4.9	4.8	4.8	4.9	5.0
		500	4.7	4.7	4.7	4.7	4.7	4.8
	{0.3, 0.3, 0.3}	100	5.0	5.2	3.9	4.0	5.9	5.2
		500	5.0	5.4	3.7	3.8	4.6	5.8
1/3-skewed	{0, 0, 0}	100	5.6	5.5	5.4	5.4	5.3	5.5
		500	4.8	4.8	4.8	4.7	4.8	4.8
	{0.3, 0.3, 0.3}	100	4.7	5.2	4.1	4.0	6.8	5.4
		500	5.8	6.4	5.0	5.1	7.2	7.5
2/3	{0, 0, 0}	100	5.2	5.2	5.1	5.1	5.0	5.2
		500	4.8	4.8	4.8	4.8	4.9	5.0
	{0.3, 0.3, 0.3}	100	6.0	6.4	4.6	4.7	**7.6**	6.6
		500	5.4	6.8	4.2	4.1	5.7	**8.3**

Note. β_1 = latent regression coefficients. PI = correctly specified partial scalar invariance model. FS = Forward stepwise specification search. AO = Alignment optimization with maximum likelihood. BAO = Bayesian alignment optimization. BAIA = Bayesian approximate invariance with alignment. BAIA-2S = two-step BAIA. Type I error rates larger than 7.5% are bolded.
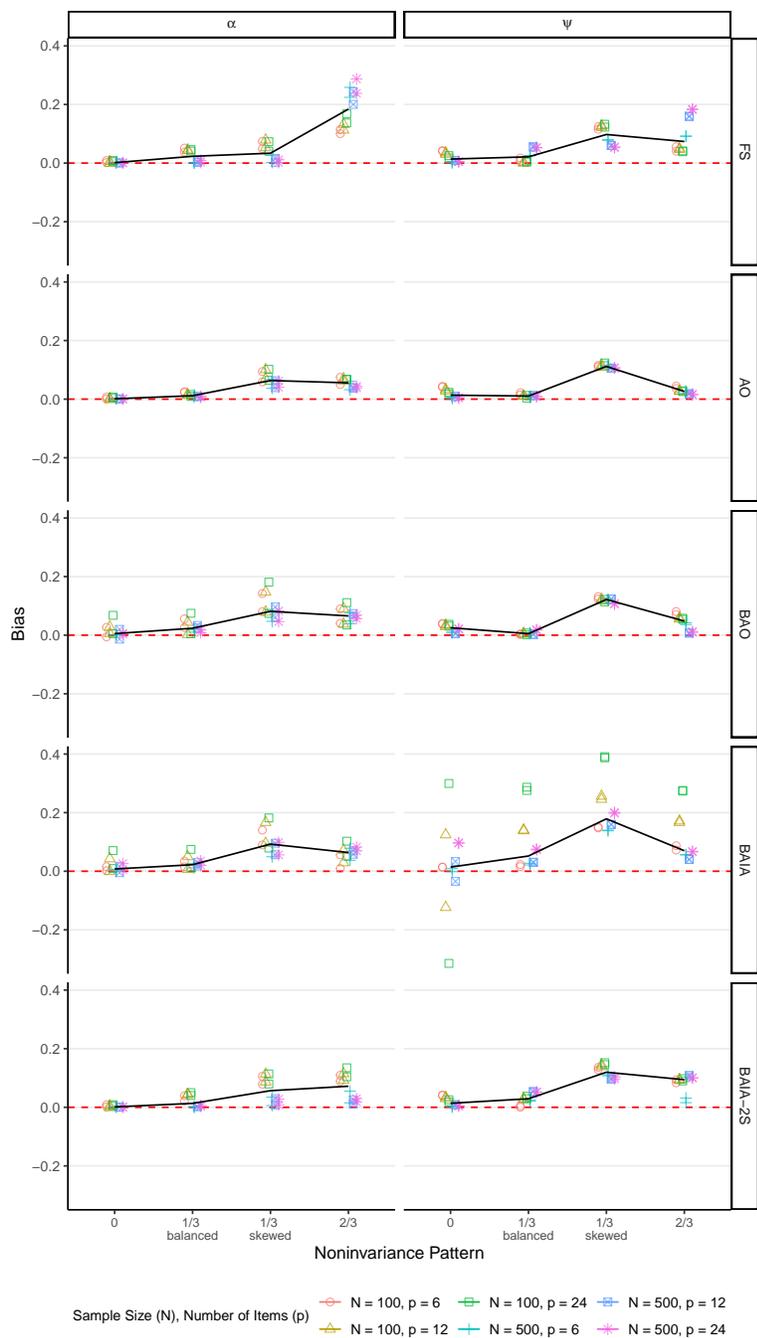
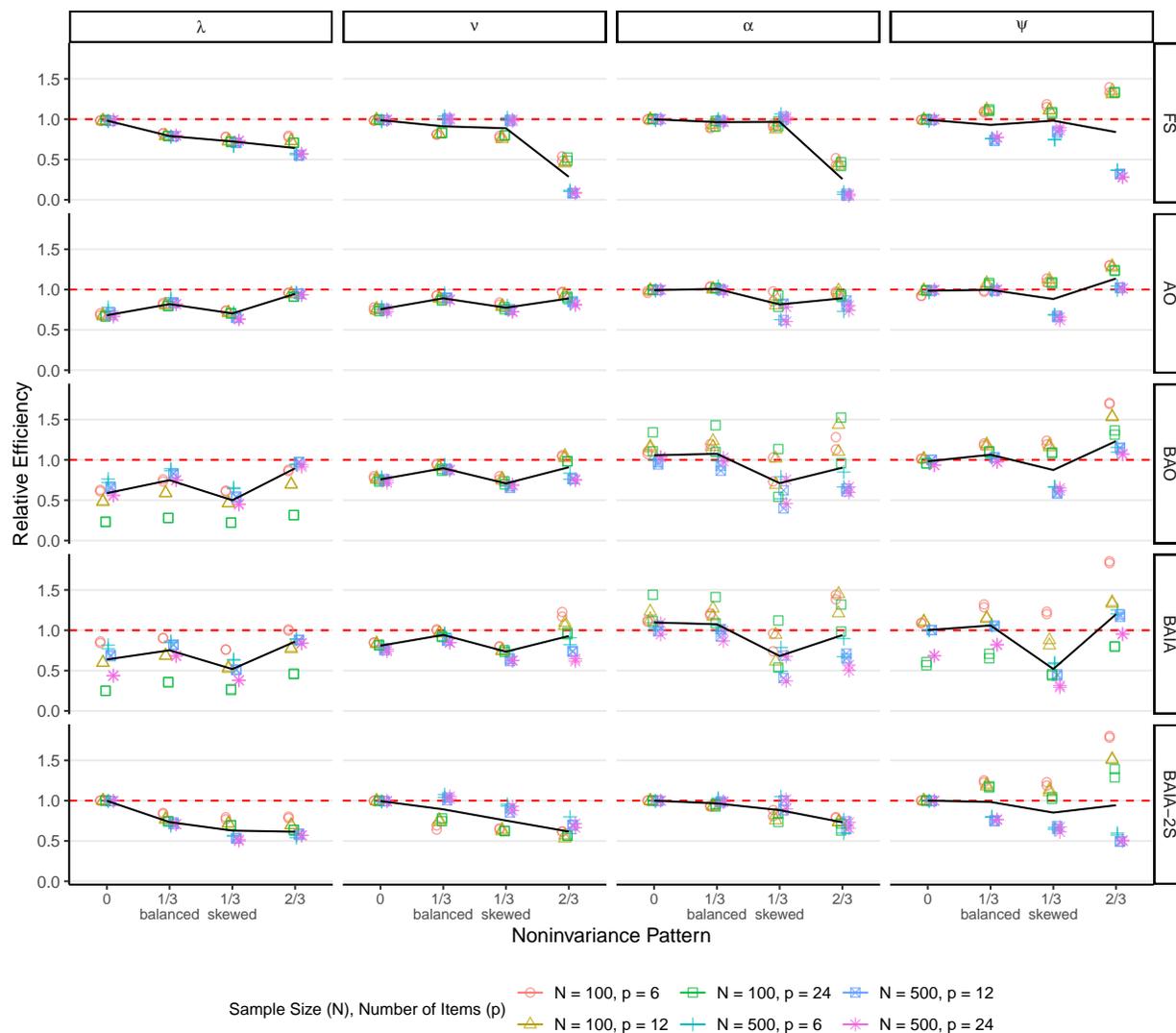*Figure 1*. Bias of Studied Estimators Relative to the Correctly Specified Partial Invariance Model in Study 1. FS = Forward stepwise specification search. AO = Alignment optimization. BAO = Bayesian alignment optimization. BAIA = Bayesian approximate invariance with alignment. BAIA-2S = two-step BAIA. α = latent mean (for Group 2). ψ = latent factor variance (for Group 2). The solid lines represent median biases across sample size and number of items conditions. The dashed lines represent bias = 0.

Figure 2. Relative Efficiency (RE) of Studied Estimators Relative to the Correctly Specified Partial Invariance Model in Study 1. FS = Forward stepwise specification search. AO = Alignment optimization. BAO = Bayesian alignment optimization. BAIA = Bayesian approximate invariance with alignment. BAIA-2S = two-step BAIA. λ = factor loading. ν = intercept. α = latent mean. ψ = latent factor variance. The solid lines represent median REs across sample size and number of items conditions. The dashed lines represent RE = 1.0, indicating that the estimator was as efficient as the correctly specified partial invariance model.
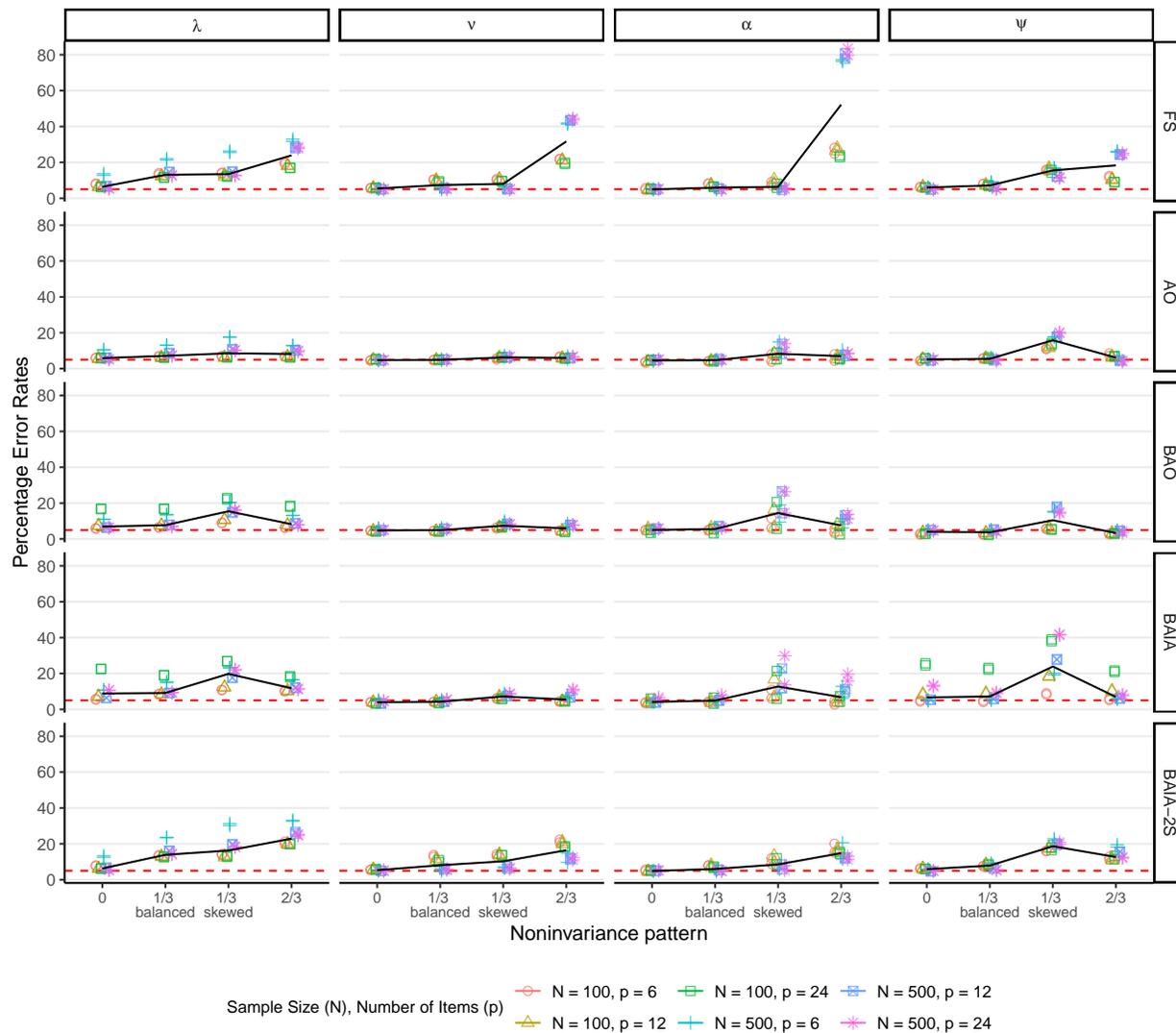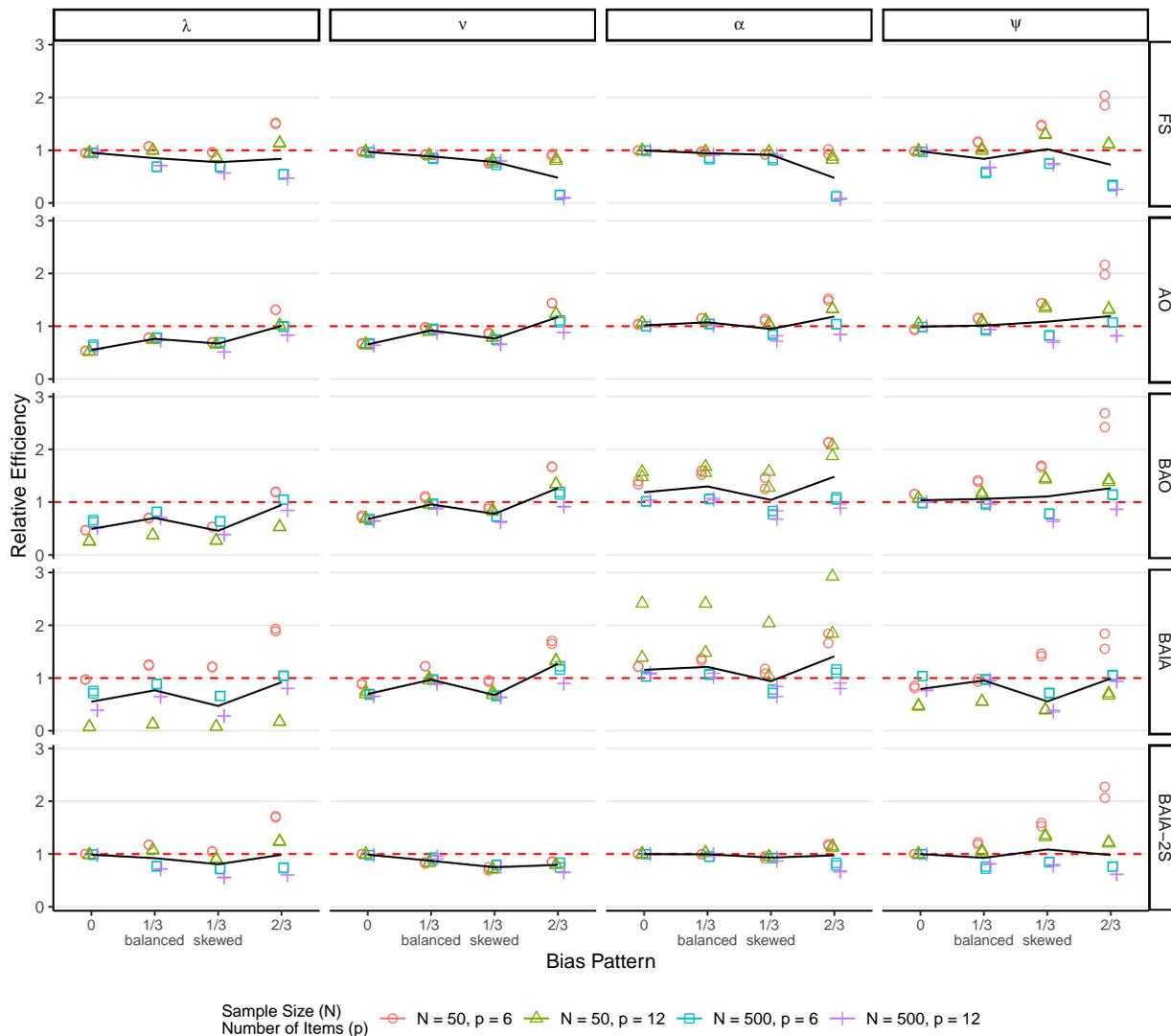
Figure 3. Percentage Error Rates of 95% Confidence (or Credible) Intervals in Study 1. FS = Forward stepwise specification search. AO = Alignment optimization. BAO = Bayesian alignment optimization. BAIA = Bayesian approximate invariance with alignment. BAIA-2S = two-step BAIA. λ = factor loading. ν = intercept. α = latent mean. ψ = latent factor variance. The solid lines represent median error rates across sample size and number of items conditions. The dashed lines represent error rate = 5%.

Figure 4. Relative Efficiency (RE) of Studied Estimators in Study 2. FS = Forward stepwise specification search. AO = Alignment optimization. BAO = Bayesian alignment optimization. BAIA = Bayesian approximate invariance with alignment. BAIA-2S = two-step BAIA. λ = factor loading. ν = intercept. α = latent mean. ψ = latent factor variance. The solid lines represent median REs across sample size and number of items conditions. The dashed lines represent RE = 1.0.
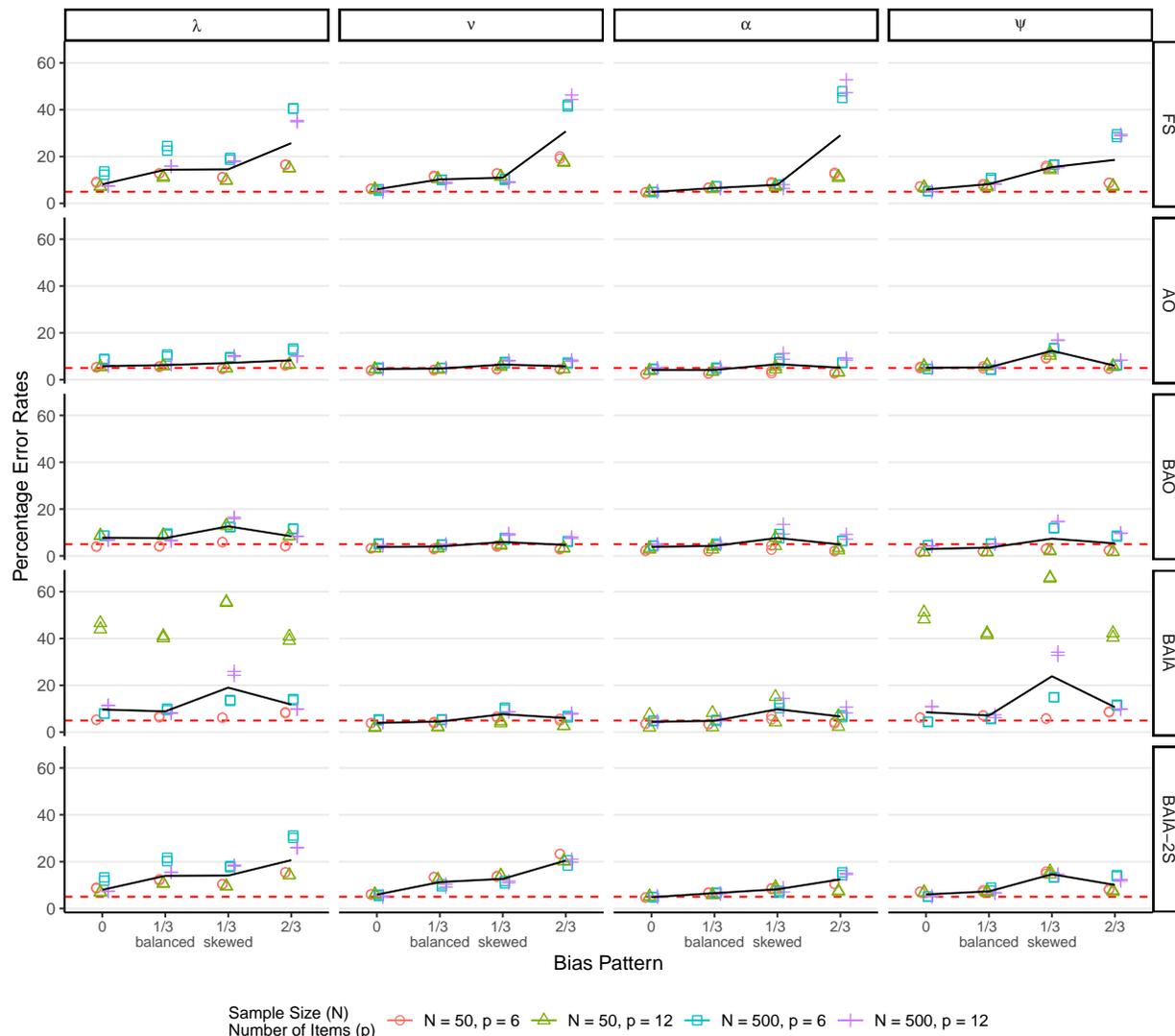
Figure 5. Percentage Error Rates of 95% Confidence (or Credible) Intervals in Study 2. FS = Forward stepwise specification search. AO = Alignment optimization. BAO = Bayesian alignment optimization. BAIA = Bayesian approximate invariance with alignment. BAIA-2S = two-step BAIA. λ = factor loading. ν = intercept. α = latent mean. ψ = latent factor variance. The solid lines represent median error rates across sample size and number of items conditions. The dashed lines represent error rate = 5%.
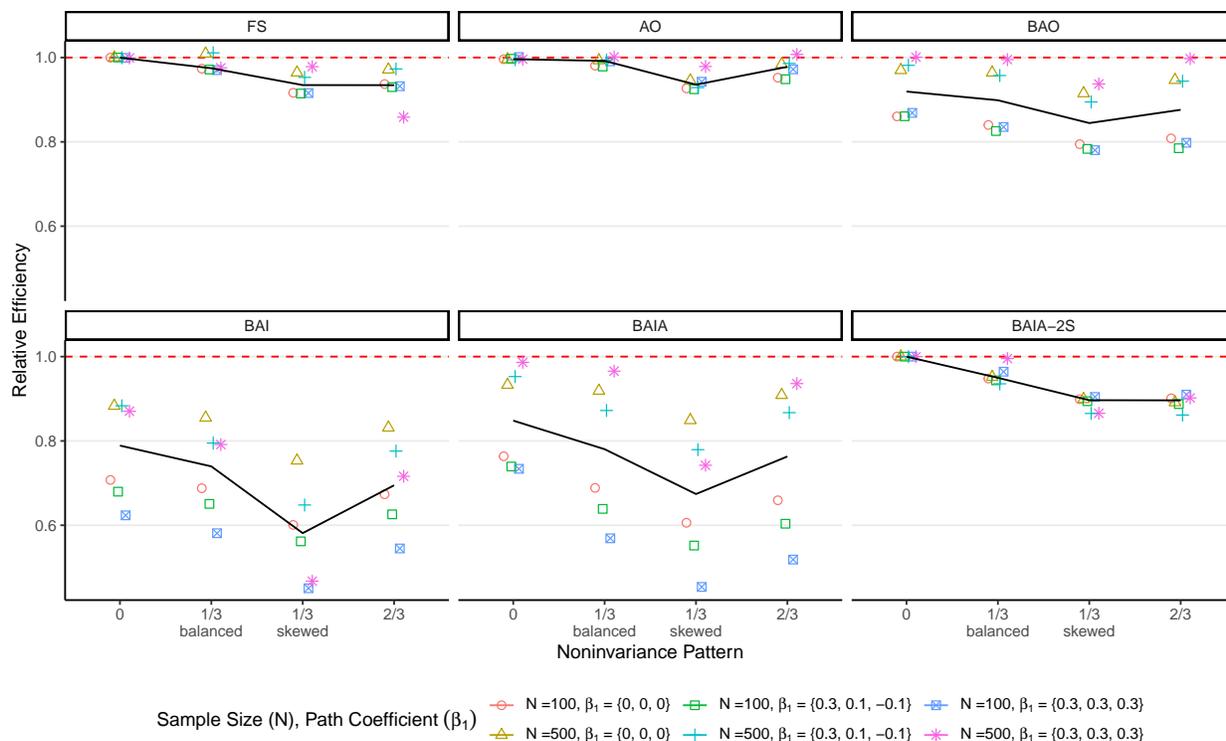
Figure 6. Relative Efficiency for Estimating Latent Regression Coefficients of the Studied Estimators in Study 3. FS = Forward stepwise specification search. AO = Alignment optimization. BAO = Bayesian alignment optimization. BAI = Bayesian approximate invariance without alignment and with small variance priors. BAIA = Bayesian approximate invariance with alignment. BAIA-2S = two-step BAIA. The solid lines represent the median bias across sample size and number of items conditions. The dashed lines represent relative efficiency = 1.0.
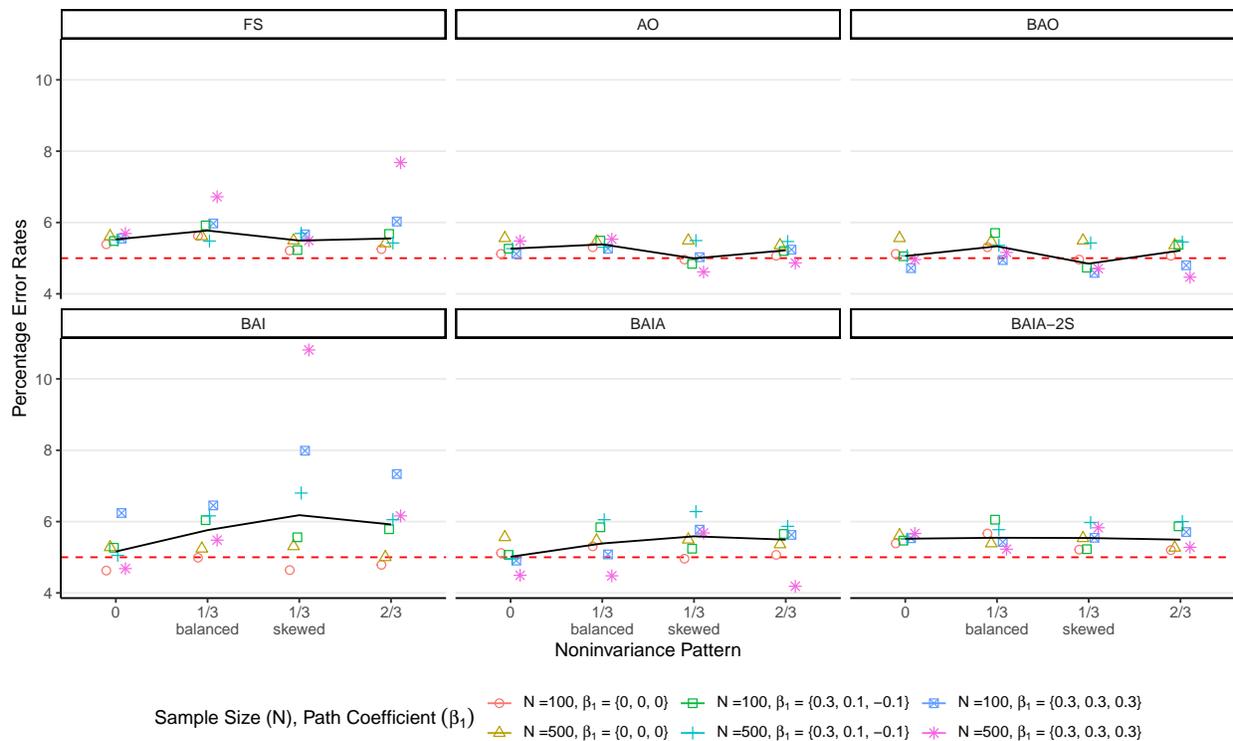
Figure 7. Error rates of 95% Confidence (Credible) Intervals for Estimating Latent Regression Coefficients of the Studied Estimators in Study 3. FS = Forward stepwise specification search. AO = Alignment optimization. BAO = Bayesian alignment optimization. BAI = Bayesian approximate invariance without alignment and with small variance priors. BAIA = Bayesian approximate invariance with alignment. BAIA-2S = two-step BAIA. The solid lines represent median biases across sample size and number of items conditions. The dashed lines represent error rate = 5%.

Appendix

Unique Factor Covariances Among Simulated Data

In Study 1, we added random unique factor covariances **W** to the indicators, where $10\mathbf{W} \sim \mathrm{LKJ}(1)$ with the shape parameter (`eta`) $= 1$. Larger values of the shape parameter pull entries in **W** closer to zero. The same realized **W** was used for all replications and conditions with same number of items. When $p = 6$, we used the R code

```
set.seed(1)
ucov <- rethinking::rlkjcorr(1, 6, eta = 1) / 10
```

which resulted in the covariance matrix

Table A1

1	2	3	4	5	6
.100	-.029	-.022	-.050	.044	-.003
-.029	.100	.047	.042	-.002	-.000
-.022	.047	.100	.063	-.021	.058
-.050	.042	.063	.100	-.071	.070
.044	-.002	-.021	-.071	.100	-.027
-.003	-.000	.058	.070	-.027	.100

The unique factor covariance matrix was similarly generated for conditions with $p = 12$ and $p = 24$, and for Studies 2 and 3 with two different seeds for random number generation.