

A Bayesian Region of Measurement Equivalence (ROME) Approach for Establishing Measurement Invariance

Yichi Zhang¹, Mark H. C. Lai¹, and Gregory J. Palardy²

¹ Department of Psychology, University of Southern California

² Graduate School of Education, University of California, Riverside

Author Note

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

Yichi Zhang  <https://orcid.org/0000-0002-4112-2106>

Files available at the Open Science Framework (<https://osf.io/e75wk/>) provide complete R and Mplus code for the examples in this manuscript. This study received support from the U. S. Army Research Institute for the Behavioral and Social Sciences (ARI) under Grant W911NF2010282 and the National Science Foundation under Grant No. 1908630. The views, opinions, and/or findings in this manuscript are those of the authors and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documents. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The content of this manuscript was presented in the online Annual Meeting of the Psychometric Society (IMPS), July 14-17, 2020.

©American Psychological Association, 2022. This is an Accepted Manuscript of an article on October 5, 2021 to be published in Psychological Methods. This paper is not the copy of record and may not exactly replicate the authoritative document published in the APA journal. The final article is available, upon publication, at: <https://doi.org/10.1037/met0000455>

Correspondence concerning this article should be addressed to Mark Lai, Department of Psychology, University of Southern California, Los Angeles, CA 90089-1061. Email: hokchiol@usc.edu

Abstract

Measurement invariance research has focused on identifying biases in test indicators measuring a latent trait across two or more groups. However, relatively little attention has been devoted to the practical implications of noninvariance. An important question is whether noninvariance in indicators or items results in differences in observed composite scores across groups. The current study introduces the Bayesian *Region of Measurement Equivalence* (ROME) as a framework for visualizing and testing the combined impact of partial invariance on the group difference in observed scores. Under the proposed framework, researchers first compute the *highest posterior density intervals* (HPDIs)—which contain the most plausible values—for the expected group difference in observed test scores over a range of latent trait levels. By comparing the HPDIs with a predetermined range of values that is practically equivalent to zero (i.e., region of measurement equivalence), researchers can determine whether a test instrument is practically invariant. The proposed ROME method can be used for both continuous indicators and ordinal items. We illustrated ROME using five items measuring mathematics-specific self-efficacy from a nationally representative sample of tenth graders. Whereas conventional invariance testing identifies a partial strict invariance model across gender, the statistically significant noninvariant items were found to have a negligible impact on the comparison of the observed scores. This empirical example demonstrates the utility of the ROME method for assessing practical significance when statistically significant item noninvariance is found.

Keywords: structural equation modeling, measurement invariance, equivalence testing, Bayesian

A Bayesian Region of Measurement Equivalence (ROME) Approach for Establishing Measurement Invariance

When using psychological test scores to compare groups or as a criterion to select individuals across groups, it is important to ascertain the test measures the same latent construct in the same way for each group (Stark et al., 2004). If individuals from different groups who have the same level of the latent construct have similar performance on the psychological test, then the test is *measurement invariant* (Meredith & Millsap, 1992). However, if the test has different measurement properties across groups for individuals with the same level of the latent construct, such as consistently giving a higher score for members of one group, it violates measurement invariance or measurement equivalence (Millsap & Kwok, 2004). In practice, it is common that only *partial measurement invariance* holds, meaning that only a subset of indicators is measurement invariant (Byrne et al., 1989; Lai et al., 2019).

There is an abundance of research on methods for examining measurement invariance. However, this literature mainly focuses on identifying noninvariant indicators, and there is a dearth of studies on the practical impact of noninvariance on group differences in the latent construct level or score. For example, when organizations or researchers use psychological tests to compare individuals or make classifications, they typically use the total test score. Thus, the interest is whether there is a difference in total scores rather than on specific indicators. To address this gap in the literature, the current study proposes a Bayesian framework for testing practical invariance in terms of the total score when one or more indicators have been flagged to be noninvariant. Compared with traditional approaches for testing measurement invariance, the proposed method allows researchers to directly quantify the degree of invariance violation in a relevant and useful way for substantive applications.

Background

Establishing measurement invariance is important because group differences in the observed test scores¹ are often difficult to interpret if invariance does not hold. For example, the Positive and Negative Affect Schedule (PANAS), a self-reported scale composed of words describing feelings and emotions (Watson et al., 1988), could function differently for participants from different cultural backgrounds. That is because the negative affect (NA) subscale of PANAS has indicators “shame” and “guilt,” whose interpretations tend to vary across cultures. Previous studies found that in collectivistic cultures, shame and guilt were viewed as slightly less negatively compared to individualistic cultures because these two words also indicated self-reflections (Eid & Diener, 2001; Sheikh, 2014). Thus, when the NA subscale of PANAS is used to assess negative affect across cultures, it is unclear whether observed differences in test scores are caused by real differences in negative affect or the different cultural meanings of test items. Therefore, without establishing measurement invariance, any research conclusions made on the NA subscale of PANAS might be inaccurate because the observed group differences in the subscale scores are confounded with measurement bias.

Traditional Approach for Testing Measurement Invariance

Two commonly used frameworks for testing measurement invariance are multiple-group confirmatory factor analysis (MG-CFA; Jöreskog, 1971) and item response theory (IRT; Millsap & Everson, 1993). Despite the different traditions and procedures of the two frameworks for invariance testing, many popular IRT models share the same underlying factor model as with MG-CFA (Takane & Leeuw, 1987). The current study focuses on the MG-CFA framework but

¹ In this article, observed scores and test scores are used interchangeably to represent the observed scores of the test instrument.

also references some IRT methods, as the proposed method for quantifying noninvariance is equally applicable to IRT models.

For simplicity, in this manuscript, we assume that the psychological test or subscale is unidimensional such that it measures only one latent construct. The MG-CFA method assumes a model where p indicators measure the underlying latent construct for K groups and has the form,

$$\mathbf{y}_{ik} = \boldsymbol{\tau}_k + \boldsymbol{\lambda}_k \eta_{ik} + \boldsymbol{\epsilon}_{ik}. \quad (1)$$

where $\mathbf{y}_{ik}$ is a $p \times 1$ vector of the observed score variables for the i th person in the k th group, $\boldsymbol{\tau}_k$ is a $p \times 1$ intercept vector, $\boldsymbol{\lambda}_k$ is a $p \times 1$ vector of factor loadings, η_{ik} is the latent score for the i th person in the k th group, and $\boldsymbol{\epsilon}_{ik}$ is the $p \times 1$ vector of unique factor variables. Also, the model assumes that $\boldsymbol{\epsilon}_{ik}$ follows a multivariate normal distribution with a mean vector of 0s and variance-covariance matrix $\boldsymbol{\Theta}_{ik}$. We further assume local independence, such that $\boldsymbol{\Theta}_{ik}$ is a diagonal matrix with elements $\theta_{1k}, \theta_{2k}, \dots, \theta_{pk}$.

Under this framework, four levels of factorial invariance have been proposed (Meredith, 1993; Millsap, 2007). The first level is configural invariance, which requires the same factor structures across groups, such as having the same number of factors and the same item patterns. The second level is weak invariance, which requires the factor loadings to be the same across groups. The third level is strong invariance, which, in addition to the factor loadings, also requires intercepts to be equal across groups. Finally, the highest level is strict invariance, which requires the factor loadings, intercepts, and the unique factor variances to be the same across groups.

Within the MG-CFA framework, there are several approaches for testing measurement invariance. Frequentist approaches include the likelihood ratio test (LRT) in the framework of traditional significance testing (Steiger et al., 1985). The LRT is a chi-square test for comparing

two nested models, one with equality constraints as implied by a particular level of invariance and the other without such constraints. This approach tests whether the parameters statistically differ across groups without showing the practical implications of such noninvariance.

Furthermore, the LRT is sensitive to sample size. As sample size increases, trivial differences among groups can generate a statistically significant result (Meade, 2010). To combat this limitation, researchers use other goodness-of-fit indices, such as *root-mean-square error of approximation* (RMSEA) and *comparative fit index* (CFI), to quantify the level of noninvariance. However, similar to LRT, these fit indices are difficult to interpret as they do not directly convey the degree of the invariance violations (Lai et al., 2019; Putnick & Bornstein, 2016). Also, researchers have proposed other indices and cutoff values, but there is a lack of agreement on which to adopt and when. These alternatives require specific research designs and thus have limited generalizability (Svetina et al., 2020).

Both IRT and goodness-of-fit indices follow the traditional null hypothesis significance testing framework (NHST), which does not allow researchers to directly support the null hypothesis of invariance. Under NHST, researchers can only reject the null hypothesis or fail to reject the null hypothesis. Failing to reject the null hypothesis that the tested parameters are equal is not equivalent to accepting it, because statistical power may be insufficient to capture the noninvariance (Kline, 2016).

Recent Development

In order to directly support measurement invariance, Yuan and Chan (2016) recently proposed the use of equivalence testing in the multigroup structural equation modeling (SEM) context. Unlike the traditional null hypothesis, equivalence testing hypothesizes that the misfit between models with invariance constraints, usually expressed in the metric of model fit indices

such as RMSEA, is greater than a prespecified amount ϵ , which can be any small positive number. In other words, instead of focusing on rejecting or failing to reject a null hypothesis of absolute equivalence in fit between models, researchers aim to provide evidence that the degree of noninvariance is less than a prespecified threshold that is considered negligible. If the null hypothesis under equivalence testing is rejected, this means the difference between models is negligible, so researchers can conclude that the invariance holds. The meaning of power also changes with the equivalence testing. In NHST, power is the probability of detecting a difference in the tested parameters given the parameters are different in the population. Whereas, in equivalence testing, power is the probability of supporting the alternative hypothesis that the difference in the tested parameters is smaller than a prespecified amount, given that the alternative hypothesis is true in the population (Yuan & Chan, 2016). Failure to reject the null hypothesis for an equivalence test does not mean insufficient evidence for differences in tested parameters anymore; it only indicates insufficient evidence to support that factorial invariance is within a tolerable size. A limitation of equivalence testing, however, is that it quantifies the impact of noninvariant indicators in the unit of model fit indices, such as RMSEA, which makes it hard to interpret the implications of noninvariance on psychological tests (Shi et al., 2019).

Previous research has also proposed methods to test measurement invariance under the Bayesian framework. The Bayesian framework regards parameters as random variables, and thus posterior distributions of estimated parameters can be obtained. One Bayesian approach, proposed by Shi et al. (2019), focuses on the cross-group differences on factor intercepts and loadings, and allows researchers to establish invariance if the posterior distributions of those differences have high probabilities of being close to zero (we discuss this in the next section).

Nonetheless, differences in factor loadings and intercepts, which are abstract model parameters, are difficult to interpret, especially in a practical sense.

Practical Significance of Noninvariance

In practice, test users or applied researchers are typically more interested in the practical implications of measurement noninvariance in observed score units than the change in model fit indices. As noted in a comprehensive review by Nye et al. (2019), around 89% of the group mean comparison studies and 58% of the predictive analyses conducted after testing for measurement invariance focused on observed scores.

Several indicator-level effect size indices have been proposed to assess the impact of measurement noninvariance on observed scores. Under the CFA framework, Nye et al. (2011) and Nye et al. (2019) proposed d_{MACS} and d_{MACS_Signed} , the signed and unsigned versions of an effect size index that reflects the observed difference for each item due to measurement noninvariance. Gunn et al. (2020) provided variations of these two indices, which they used different standard deviations to standardize the impact of noninvariance. In addition, they proposed two other similar indices that can be used with single factor analysis.

In addition to the indicator scores, total test scores are often of interest for selections or classifications. For example, consider the Center for Epidemiological Studies-Depression (CES-D; Radloff, 1977), a clinical screening test for assessing depressive symptoms that uses a cutoff score of 16 for mild to moderate depression. While there is an abundance of research examining measurement invariance of the CES-D across cultural groups (e.g., Kim et al., 2009; Yang et al., 2015), most studies have only identified noninvariant indicators without examining the impacts of noninvariance on total test scores. Such information, however, is of limited utility for decisions about whether the scale itself is biased in making clinical diagnoses (e.g., Lai et al.,

2017; Millsap & Kwok, 2004). For that essential decision, it is necessary to test measurement invariance and investigate its implications on total test scores.

Although there are some discussions on indicator-level effect size indices in the literature, there has been little discussion on the effects of indicator noninvariance on total scores. One index related to the current research is the DTFR statistic proposed by Stark et al. (2004; see also Chalmers et al., 2016) under the IRT framework, examining the expected group mean difference in total test scores. They also proposed d_{DTF} , an effect size index that divides DTFR by the standard deviation of the focal group's scores (Stark et al., 2004; see a similar index in Meade, 2010). A highly similar index to DTFR, Δ_{mean} , was later discussed in Nye & Drasgow (2011) under MG-CFA. However, a limitation of DTFR and Δ_{mean} is that they show the degree of measurement invariance violations by a single number representing the average bias across levels of the latent construct. Therefore, if the bias favors one group on the high end of the latent construct but favors the other group on the low end, these two indices may indicate the test instrument is nearly unbiased when it is not.

In addition, Chalmers et al. (2016) proposed $sDTF_0$, an index showing test bias at a given latent trait level, and developed a simulation-based procedure for obtaining approximate confidence intervals for $sDTF_0$. However, both DTFR and $sDTF_0$ are developed in the IRT framework, and similar indices have not been developed in the CFA framework. Furthermore, previous research has not provided inferential procedures for using these effect size indices to *support* measurement invariance, which is the focus of the current study.

Focus of the Present Study

To address the limitations of current measurement invariance practices, we propose the *region of measurement equivalence* (ROME) method for directly supporting measurement

invariance and testing its impact on total scores using the Bayesian structural equation modeling (BSEM) framework. The ROME method is similar to equivalence testing, but it focuses on a metric that is easier to interpret, namely, the impact of biased indicators on group differences in total test scores.² It should be pointed out that the ROME method complements, but does not replace, conventional methods of identifying noninvariant indicators or quantifying bias on the metric of model parameters, as there are conditions researchers need to obtain such information. Using ROME for invariance research provides more detailed information on how noninvariance affects the total scores, which is usually of interest when using tests for selecting individuals or comparing groups.

Measurement Invariance Testing via Bayesian ROME

The Bayesian Framework

The Bayesian framework has been increasingly used in *structural equation modeling* (SEM) over the past few years (Lee, 2007; Hoyle, 2012; Shi et al., 2019). Assume η is the *latent variable*, which is the unobservable attribute that we are interested in, y is the observable value of η , and θ is one of the unknown parameters of SEM (Hoyle, 2012). Our goal is to determine $p(\theta|y)$, the *posterior distribution* of parameter θ given the observed data y . According to the *Bayes' theorem*, $p(\theta|y) \propto p(y|\theta)p(\theta)$, where $p(y|\theta)$ represents the *likelihood* of θ given y , and $p(\theta)$ stands for the *prior distribution* of θ . The theorem suggests that the posterior distribution of a parameter is proportional to the product of the likelihood of the parameter given the observed value and prior distribution of that parameter.

As stated above, the posterior distribution of a parameter can be acquired through weighting the prior distribution of the parameter by data. In the traditional frequentist view,

² Here total test scores represent the sum of observed scores or test scores of all items.

parameters are treated as constants, whereas in the Bayesian framework, parameters are regarded as random variables with probability distributions (Song & Lee, 2012). Thus, summary statistics, such as mean and variance, can be calculated from the posterior distribution (Hoyle, 2012).

However, when the model involves latent variables such as the CFA model, it can be computationally demanding to get the mean or variance because it requires numerical integration. Under this situation, *Markov Chain Monte Carlo* (MCMC), a random sampling method that draws samples repeatedly from the posterior distribution, can summarize the distribution and obtain useful information such as mean and variance (Gill, 2008). This allows MCMC to provide estimates while avoiding the potential computational challenges noted for numerical integration.

Highest Posterior Density Interval (HPDI)

In the Bayesian framework, because each parameter is assumed to have a probability distribution, we can obtain the *credible interval*, a fixed interval in which the parameter value falls with a certain probability (Box & Tiao, 1993). For example, a 95% credible interval means, integrating information from the prior distribution and the observed data, there is a 95% probability that the parameter is inside the credible interval (Kruschke, 2014), which is different from a frequentist confidence interval. One type of credible interval of particular interest is the *highest (posterior) density interval* (HPDI), or HDI, inside which any point has a higher probability than points outside that interval. The formal definition of HPDI is given by Kaplan (2014, p. 96):

Let $p(\theta|y)$ be the posterior probability density function. A region R of the parameter space θ is called the $100(1 - \alpha)\%$ HPDI if

$$1. p(\theta \in R|y) = 1 - \alpha$$

2. For $\theta_1 \in R$ and $\theta_2 \notin R$, $P(\theta_1|y) \geq P(\theta_2|y)$

In other words, an HPDI essentially represents the degree of uncertainty about a parameter (Kruschke, 2018). When the sample size is large and/or with informative priors, the posterior distribution is likely sharp and the HPDI is narrow, which reflects a high degree of certainty about the parameter estimates. In contrast, if the posterior distribution is likely flat and the HPDI is wide, the parameter estimates have high degrees of uncertainty.

Some previous research has used HPDI for statistical testing by comparing HPDI with a *region of practical equivalence* (ROPE) (Kruschke, 2014; Shi et al., 2019), as discussed below.

ROPE

For some applied studies, researchers are interested in estimating the difference in parameters across groups and whether these differences are negligible. Kruschke (2014) proposed using ROPE for this purpose. Specifically, ROPE represents a range of parameter values that are practically equivalent to the null value (Kruschke, 2014).

Comparing an HPDI to a ROPE can be used as a decision rule for null hypothesis testing in the Bayesian framework (Kruschke, 2018). Specifically, if the 95% HPDI is completely within the preset ROPE, there is at least a 95% chance that the parameter is practically equivalent to the null value, and, therefore, the null hypothesis should be accepted in a practical sense. On the contrary, if the 95% HPDI is entirely outside the preset ROPE, that indicates there is at least a 95% chance that the parameter is practically different from the null value, and, therefore, the null hypothesis is rejected. When the 95% HPDI is neither completely inside the ROPE nor completely outside the ROPE, whether the null hypothesis should be rejected remains inconclusive because some of the most credible values are practically equivalent to the null value, but others are not.

The HPDI and ROPE decision rule requires specifying the limits of the ROPE. Since there are no substantive guidelines on how to decide the appropriate range of values that are considered practical equivalence, Kruschke (2014) recommended using a range from -0.1 to 0.1 of a standardized parameter, so that the region contains values of standardized mean difference effect size less than or equal to 0.1 (Cohen, 1988). However, this cutoff value is arbitrary, and researchers should decide the ROPE based on a rationale for gauging the practical differences in total scores.

In addition to the limits of ROPE, the width of an HPDI is also an influencing factor of the decision. When an HPDI has a narrower range reflecting a high degree of certainty, it is more likely to conclude the parameter is practically equal to the specified null value or substantially different from the null value. However, if the HPDI is wide, the HPDI limits may overlap with the ROPE and the zero point, leading to inconclusive results (Kruschke, 2018).

Region of Measurement Equivalence (ROME)

The current study applies the concept of ROPE to measurement invariance analysis, which we call ROME. Specifically, a ROME is a range of values on total test scores that are practically equivalent to no bias across groups. Similar to ROPE, researchers could use ROME and HPDI to make decisions for null hypothesis testing in the Bayesian framework and evaluate measurement equivalence of the scale.

Computing the Expected Difference in Total Scores

In multiple group studies, researchers are usually interested in group differences and whether these differences are negligible. The current study investigates the impact of measurement noninvariant indicators on observed group differences by calculating the expected difference in total test scores across a range of latent trait values. Based on the multi-group

confirmatory analysis model in equation (1), the expected score of the j th indicator for the i th person with given η in the k th group, could be written as:

$$E(y_{ijk}|\eta) = \tau_{jk} + \lambda_{jk}\eta. \quad (2)$$

This can be extended to represent the expected total test score, T_{ik} , for the i th person with a given η in the k th group. As shown on the right-hand side of equation 3, the expected sum of y_{ijk} of p indicators, can be further written as the sum of τ_{jk} of p indicators plus the sum of λ_{jk} of p indicators times the latent score η .

$$E(T_{ik}|\eta) = E(y_{i1k} + y_{i2k} + \dots + y_{ipk}|\eta) = \sum_{j=1}^p \tau_{jk} + \eta \sum_{j=1}^p \lambda_{jk} \quad (3)$$

The expected difference in total test scores between two groups for people with the same latent score η is:

$$E(D_{2-1}|\eta) = E(T_2|\eta) - E(T_1|\eta). \quad (4)$$

Note that when strict invariance holds, D_{2-1} equals zero since two people with the same latent score are expected to have the same observed total score. Therefore, the larger the $E(D_{2-1}|\eta)$, the larger the impact noninvariance has on the total test score.

When the MG-CFA model is fitted as a BSEM model, one can obtain the posterior distribution for each parameter, which can be used to get the 95% HPDI for expected total test scores of each group, as well as the group difference in the expected total scores. We argue that the consequence of partial invariance on total test scores is often more practically useful than detecting measurement invariance on specific indicators. Thus, the ROME method estimates

how much the expected total score differs between two groups across latent trait levels (η), which is a practical statistic of the impact of partial invariance.³

Note that in a factor model, η is unbounded and can take on any value on the real line, and it can be shown that, unless the loadings are equal across groups, $E(D_{2-1}|\eta)$ will be outside of any finite ROME for some values of η . Therefore, we only consider a range of η that captures a typical range of the population. In this study, we choose $\pm 2SD$ as the range of η , which captures about 95% of the population under a normal distribution. However, if researchers are interested in using the test for people with higher or lower latent trait levels, $E(D_{2-1}|\eta)$ can be computed on a different range of η .

Presetting the ROME

The Bayesian framework enables researchers to use the 95% HPDI and ROME to establish measurement invariance. As a starting point, we set the ROME following the convention suggested by Kruschke (2014), which is $[-0.1s_p, 0.1s_p]$. Specifically, s_p in the proposed method is defined as the pooled standard deviation of the total test scores of the psychological test being evaluated:

$$s_p = \sqrt{\frac{(n_k - 1)s_k^2 + (n_j - 1)s_j^2}{n_k + n_j - 2}} \quad (5)$$

³ Although the proposed method shares some commonalities with the $sDTF_\theta$ statistics proposed by Chalmers et al. (2016), these two methods differ in three major aspects. First, $sDTF_\theta$ was developed under the NHST framework, which indicates its goal is to reject invariance. In contrast, ROME is proposed to directly support practical invariance, which compares the impact of noninvariance with a range of negligible group differences. Second, $sDTF_\theta$ is developed under the IRT framework with binary and categorical items. In contrast, ROME is proposed under MG-CFA and can be used with continuous indicators and categorical indicators. We illustrate ROME with continuous indicators in the empirical example and provide a parallel example using categorical indicators in the supplemental materials (<https://osf.io/e75wk/>). Third, multivariate normality is an essential assumption of the Monte Carlo method used by $sDTF_\theta$, but it is not required for MCMC used by ROME. When the sample size is small, MCMC can give more accurate inferences than the Monte Carlo method (Lee & Song, 2004).

In this case, n_k and s_k represent the number of observations and the sample standard deviation of total test score for the k th group, whereas n_j and s_j represent the number of observations and the standard deviation of sample total test score for the j th group.

After setting the ROME for group differences and obtaining the 95% HPDI for the expected difference in total scores, researchers can examine whether measurement invariance holds in a practical sense, and then quantify the impact of group differences on total scores. However, the proposed method is only applicable when a partial invariance model is identified, meaning that at least one indicator is found or allowed to be noninvariant. A summary of the steps for applying the proposed ROME method is discussed below, followed by an illustrative empirical example.

Steps for Using ROME to Test for Measurement Invariance

Following conventional measurement invariance testing,⁴ assume that at least one indicator was found noninvariant. The Bayesian ROME can be implemented using the following steps:

1. Set the ROME, a range of values where the group differences in the expected total test scores given η is practically equivalent to zero.
2. Fit Bayesian MG-CFA model to data and obtain posterior distributions of the model parameters for each group (intercepts, factor loadings, and uniqueness).
3. Compute the posterior distribution and the 95% HPDI for the expected difference in total scores between groups, given η .

⁴ Researchers can use either the frequentist or the Bayesian procedures to identify noninvariant indicators or items, and use ROME to evaluate practical invariance. We used frequentist CFA to identify noninvariant indicators in the below example because it is one of the more popular approaches to test MI. However, researchers can use Bayesian procedures such as the Bayes factor (Kruschke, 2011) or the deviance information criterion (Verhagen & Fox, 2012).

4. Compare the 95% HPDI in (3) with the preset ROME to evaluate whether measurement invariance holds practically at the observed score level.

Empirical Example

Here we use R to demonstrate the proposed method for testing measurement invariance with continuous CFA as an approximation of the discrete items for simplicity, and in the supplemental materials, we demonstrate using ROME with Mplus and R for item factor analysis (Wirth & Edwards, 2007) with ordinal items. The R codes and Mplus syntax for the example can be found at the Open Science Framework (OSF; <https://osf.io/e75wk/>). This example uses a nationally representative sample of 10th graders from the Educational Longitudinal Study of 2002 (ELS: 2002), collected by the National Center for Education Statistics (U.S. Department of Education, 2004). ELS: 2002 follows students through high school to postsecondary years and includes extensive student survey items and items from parents, teachers, and school administrators.

The current study investigated measurement invariance of a 5-item scale of math-specific self-efficacy across gender. Self-efficacy, which was described by Bandura (1997) as the “beliefs in one’s capabilities to organize and execute the courses of action required to manage prospective situations” (p. 2), is one of the most widely studied dispositions in psychology and education (Palardy, 2019). The scale assessed participants’ math-specific tasks or skills and consisted of five items rated on a 4-point Likert-type scale (1 = *Almost never*, 4 = *Almost always*). Item descriptions are provided in Table 1. As shown in Table 1, all inter-item correlations exceed 0.68. All participants were included in the analysis except those that did not respond to any items ($N = 11,663$; 47.69% male, 52.31% female).

If this scale is used to assess students' self-efficacy in math and compare their scores across gender groups, it is important to ensure the scale works the same for both genders. Thus, this example's objective is to test whether the scale of math self-efficacy is measurement invariant across gender. We first describe the results from conventional invariance testing and then illustrate the ROME method using the steps proposed above, which also shows the impact of noninvariant items on the expected difference in total scores between genders in plots.

Conventional Invariance Testing with MG-CFA

A series of multiple-group factor models were fit to find the baseline factorial invariance model. Table 2 provides the fit indices for each model and the Appendix provides additional details. We used the *lavaan* package (v0.6-5; Rosseel, 2012) in R to conduct the measurement invariance analysis. Full information maximum likelihood estimation was used for the CFA models. Parallel analysis suggested that one component, with an eigenvalue of 4.05 and an explained variance of 80.96%, should be retained. We tested the one-factor model of math self-efficacy described in Schaefer (2009), which allowed the unique factor covariances between items 1 and 2 and 2 and 3 to be correlated. The initial model had a good fit (RMSEA = .056, 90% CI [0.048, 0.065], CFI = .998), so we further tested the configural invariance and weak invariance. Results indicated that item 5 has differential loadings for males and females. Consequently, a partial strong invariance model that constrained the factor loadings for all items except item 5 and intercepts for all items was fit to data. The results showed items 2, 4, and 5 were noninvariant in their intercepts. Therefore, the intercepts of these items were freed to improve model fit.

Lastly, a partial strict invariance model, which freed the factor loading for item 5 and intercepts for items 2, 4, 5, was fit to the data. Since the chi-square difference test showed a

statistically significant difference from the partial strong invariance model, we further relaxed the constraint in unique factor variance of items 4 and 5 so that the model had an improved fit. In addition, the unique factor covariances between items 1 and 2, 2 and 3 were tested.⁵ We found evidence that the covariance between items 1 and 2 was noninvariant across groups. To conclude, items 1 and 3 showed strict invariance, items 2 and 4 showed weak invariance, and item 5 was noninvariant.

Step 1: Set ROME for Group Differences

In this illustration, we follow the convention of using a ROPE limit of $\pm 0.1s_p$, the pooled standard deviation of the total test scores. Given that $s_p = 4.12$, our ROME was set as $[-0.412, 0.412]$.

Step 2: Obtain Parameter Posterior Distributions

Next, we fitted the partial strict invariance model obtained from the conventional frequentist CFA using BSEM with the *blavaan* (v. 0.3-8; Merkle & Rosseel, 2018) R package. The default weakly informative priors of *blavaan* were used for all estimated parameters. Specifically,

$$\lambda \sim N(0, 10)$$

$$\tau \sim N(0, 32)$$

$$\theta^{1/2} \sim \text{Gamma}(1, 0.5)$$

$$\alpha \sim N(0, 10)$$

$$\psi^{1/2} \sim \text{Gamma}(1, 0.5),$$

where α and ψ are the latent means and latent variances for the female group. Note that in *blavaan*, the normal prior is parameterized using the mean and the standard deviation, and for

⁵ See Byrne (2004) for more discussion on testing unique factor covariances.

scale parameters, the priors are set on the standard deviation (i.e., $\theta^{1/2}$ and $\psi^{1/2}$). For this BSEM model, we set the number of chains to three, the number of warmup iterations per chain to 500, and the number of post-warmup iterations per chain to 1,000. The target acceptance rate was 95% (`adapt_delta = 0.95`). To assess convergence, we ensured that there were no divergent transitions post-warmup and that $\hat{R} < 1.01$, effective sample sizes > 800 for all parameters (Vehtari et al., 2020). The trace plots of the factor loadings for each item are provided in Figure 1. The figure indicates that the factor loadings converged rapidly, and three chains mixed well with each other.

The parameter estimates extracted from *blavaan* are shown in Table 3. Our reference group was female, so the latent factor mean for female was fixed to 0, and the latent factor variance for female was set to 1. In contrast, the male group had a mean of 0.323 (posterior SD = 0.021) and a variance of 1.020 (posterior SD = 0.030). The result suggested that males have higher latent self-efficacy in math than females, which is consistent with the research literature (Huang, 2013).

Step 3: Compute the Expected Difference in Total Scores Given η and its 95% HPDI

We then obtained posterior distributions for estimated parameters and the 95% HPDI for the expected group difference in total scores. Specifically, for each level of η in the range of two SD below and two SD above the mean of the female sample on η , we calculated the expected difference in total test scores across gender using female score minus male score, and generated a 95% HPDI. Here the mean expected differences between female and male for $M \pm 2SD$ of math-specific self-efficacy were reported because most students were within this range. The results indicated that for students with math self-efficacy equal to the sample mean for the female group (i.e., $\eta = 0$), the posterior mean group difference in total scores between female and male was

0.102, and the 95% HPDI was [0.057, 0.146]. For students who are two standard deviations above the mean math-specific self-efficacy (i.e., $\eta = 2$), the posterior mean group difference in total scores was 0.172, and the 95% HPDI was [0.113, 0.232]. For students who are two standard deviations below the mean math-specific self-efficacy (i.e., $\eta = -2$), the posterior mean group difference in total scores was 0.032, and the 95% HPDI was [-0.031, 0.095]. Thus, for students with the same latent level of self-efficacy in math, females tended to report a higher observed self-efficacy score than males by less than 0.2 point on a range of 15 points (from the lowest possible score of 5 points to the highest possible score of 20 points).

Step 4: Compare the 95% HPDI with ROME and Determine Measurement Invariance

As shown in Figure 2, the expected total score for females was higher than the expected total score for males given the same level of self-efficacy in math from $\eta = -2$ SD to $\eta = 2$ SD. For students who have a self-efficacy score of approximately two standard deviations above the mean, the 95% HPDI shows that the expected gender difference is between 0.113 and 0.232 points. The expected group difference in total scores is shown in Figure 3. The graph suggests that although there is a slight difference in total scores between females and males given the same trait level of math self-efficacy on this scale, it is practically negligible because the mean difference is completely within the preset ROME [-0.412, 0.412]. Stated another way, for females and males with the same level of self-efficacy, the mean group difference is not large enough to be of practical significance, although females tended to report a slightly higher score than males on this scale. Thus, we conclude, for practical purposes of using the total test scores, the math-specific self-efficacy scale is measurement invariant across gender.

Discussions

In this paper, we introduce the Bayesian ROME approach for establishing measurement invariance. It is designed to complement conventional methods of testing measurement invariance by quantifying the impact of noninvariance on the total test score. It provides the expected difference in total scores among groups with a 95% HPDI. By comparing the 95% HPDI with the preset ROME—a region of practical measurement equivalence, researchers can establish measurement invariance of a test instrument and decide whether the group difference caused by biased indicators is negligible.

Compared to the traditional frequentist approach, the ROME approach can provide more information. First, it can directly support the hypothesis of measurement invariance, whereas traditional frequentist methods either reject or fail to reject this hypothesis. The Bayesian parameter estimation framework enables researchers to make statistical inferences by using 95% HPDI as a tool for hypothesis testing (Kruschke, 2018; Shi et al., 2019). Different from conventional hypothesis testing, where failure to reject the null hypothesis of measurement invariance does not correspond to accepting the null hypothesis (Cohen, 1994), the ROME approach enables researchers to confidently conclude that the measurement invariance is established by comparing the 95% HPDI with the ROME (Shi et al., 2019).

Second, the ROME approach can help researchers quantify test bias across groups in an understandable way. Previous approaches mainly focused on obtaining an all-or-none decision when testing measurement invariance. Even when effect size measures are used, they typically report the differences in factor loadings or intercepts for noninvariant indicators, which may not be intuitive for test users and applied researchers. The ROME method overcomes this shortcoming by reflecting the impact of biased indicators on the unit of the observed score. This eases the interpretations of the measurement invariance results and focuses on practical

implications of the noninvariant indicators. Moreover, the ROME approach can visualize the impact of noninvariant indicators on total test scores. If the expected group difference in total test scores is a constant, then the test instrument is biased towards one group for people who have the same level of the latent construct, and the bias is the same magnitude for everyone. However, if the expected group difference in total scores is a straight line with a nonzero slope, the noninvariant indicators have different factor loadings across groups. Consequently, the test instrument might be measurement invariant for some participants but not for others. Thus, compared to traditional approaches, the ROME method could provide additional valuable information on when and to whom the test instrument is practically invariant.

Alternative Methods for Presetting ROME

In the real data example, we set our ROME following the ROPE convention suggested by Kruschke (2014), which is to use 0.1 pooled standard deviation of the observed total score as the range of practical equivalence. However, the limit of ROME could also be set using other rules. One option is to set the ROME based on the minimal meaningful scoring unit. Any score difference less than half of this unit could be treated as negligible, which suggests ROME to be $[-0.5 \text{ minimal meaningful scoring unit}, 0.5 \text{ minimal meaningful scoring unit}]$. For example, some test instruments, like the Scholastic Assessment Test (SAT), uses 10 points as the smallest unit to distinguish people. Thus, any value within the range of 5 points can be considered practically equivalent.

Another approach of setting ROME depends on the effect size. Similar to the idea of ROPE, Lakens (2017) proposed to specify an equivalence bound such that any result that falls within it is equivalent to the absence of an effect that is practically meaningful. He suggested setting the equivalence bound to the smallest effect size of interest (SESOI), the minimum

difference between groups that can be considered statistically significant. The value of SESOI can be obtained by conducting a power analysis based on the expected effect size and is determined exclusively by sample size (Lakens, 2017). However, this approach is controversial because other researchers believe ROPE should not be set based on measurement precisions determined by sample sizes (Kruschke, 2018). In general, setting ROME limits is similar to choosing threshold values for other decision statistics such as p -values and Cohen's d . Thus, researchers should focus on the purpose of the study and think about the range of values that can be considered practically equivalent in scales that are frequently used in their fields.

Limitations and Future Directions

It should be noted that there are two prerequisites for using the ROME method. One is to have a good model fit for the partial invariance model identified by the traditional frequentist method. Having a misfit in the partial invariance model would result in an unsatisfactory fit in the BSEM model, which could produce an inaccurate 95% HPDI and further influence the conclusion about measurement invariance of the test instrument. The other prerequisite is that the ROME method only applies when partial invariance is detected by the conventional measurement invariance testing method. If the conventional Likelihood Ratio Test does not detect any violation of measurement invariance, then the two groups would have the same expected total scores so that the 95% HPDI would always be inside the ROME. In other words, it is meaningless to use the ROME method when there are no violations of measurement invariance.

A limitation of the illustrative example is that most items fail scalar invariance, which might be a concern for interpreting the latent traits as on the same scale. Although some research such as Shi et al. (2017), showed in simulation studies that a partial invariance model with one

invariant item as the anchor item was sufficient for setting the zero point for the latent trait, this result is based on the assumption that the invariant item was correctly identified. Thus, researchers need to be cautious interpreting the results of the illustrative example and future studies with partially invariant scales.

The current study can be extended to three or more groups. Extending the ROME method to more than two groups could help researchers establish measurement invariance on psychological tests that are widely used to compare multiple groups, such as surveys in multinational studies. One possible direction is to apply the proposed method with the same ROME to every pair of groups, and support invariance when all HPDIs on the pairwise difference on the expected total scores are within the preset ROME.

In summary, the current study introduced a ROME approach for testing measurement invariance. Besides identifying noninvariant indicators, the ROME method allows researchers to establish measurement invariance and determine whether the group differences caused by item biases are practically meaningful. It also provides more information, such as when and to whom the test is noninvariant, as well as graphical tools to visually show the impact of noninvariant indicators on total scores, which could be useful for research designs and communication.

References

- Bandura, A. (1997). *Self-Efficacy: The exercise of control*. W. H. Freeman.
- Box, G. E. P., & Tiao, G. C. (1993). *Bayesian inference in statistical analysis*. Wiley.
- Byrne, B. M. (2004). Testing for multigroup invariance using AMOS graphics: A road less traveled. *Structural Equation Modeling: A Multidisciplinary Journal*, 11(2), 272–300.
https://doi.org/10.1207/s15328007sem1102_8
- 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., Counsell, A., & Flora, D. B. (2016). It might not make a big DIF: Improved differential test functioning statistics that account for sampling variability. *Educational and Psychological Measurement*, 76(1), 114–140.
<https://doi.org/10.1177/0013164415584576>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). L. Erlbaum Associates.
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49(12), 997–1003. <https://doi.org/10.1037/0003-066X.49.12.997>
- Eid, M., & Diener, E. (2001). Norms for experiencing emotions in different cultures: Inter- and intranational differences. *Journal of Personality and Social Psychology*, 81(5), 869–885.
<https://doi.org/10.1037/0022-3514.81.5.869>
- Gill, J. (2008). *Bayesian methods: a social and behavioral sciences approach* (2nd ed.). Chapman & Hall/CRC.

- Gunn, H. J., Grimm, K. J., & Edwards, M. C. (2020). Evaluation of six effect size measures of measurement non-invariance for continuous outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, 27(4), 503–514.
<https://doi.org/10.1080/10705511.2019.1689507>
- Hoyle, R. H. (2012). *Handbook of structural equation modeling*. Guilford Press.
- Huang, C. (2013). Gender differences in academic self-efficacy: a meta-analysis. *European Journal of Psychology of Education*, 28(1), 1-35.
- Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, 36(4), 409–426. <https://doi.org/10.1007/BF02291366>
- Kaplan, D. (2014). *Bayesian statistics for the social sciences*. Guilford Press.
- Kim, G., Chiriboga, D. A., & Jang, Y. (2009). Cultural equivalence in depressive symptoms in older white, black, and Mexican-American adults. *Journal of the American Geriatrics Society*, 57(5), 790–796. <https://doi.org/10.1111/j.1532-5415.2009.02188.x>
- Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.
- Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, 6(3), 299–312.
<https://doi.org/10.1177/1745691611406925>
- Kruschke, J. K. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (Second edition). Academic Press.

- Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, 1(2), 270–280.
<https://doi.org/10.1177/2515245918771304>
- Lai, M. H. C., Kwok, O.-m., Yoon, M., & Hsiao, Y.-Y. (2017). Understanding the impact of partial factorial invariance on selection accuracy: An R script. *Structural Equation Modeling*, 24(5), 783–799. <https://doi.org/10.1080/10705511.2017.1318703>
- Lai, M. H. C., Richardson, G. B., & Mak, H. W. (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>
- Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4), 355–362.
<https://doi.org/10.1177/1948550617697177>
- Lee, S.-Y. (2007). *Structural equation modelling: A Bayesian approach*. John Wiley & Sons Ltd. <https://doi.org/10.1002/9780470024737>
- Lee, S.-Y., & Song, X.-Y. (2004). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavioral Research*, 39(4), 653–686. https://doi.org/10.1207/s15327906mbr3904_4
- Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology*, 95(4), 728–743.
<https://doi.org/10.1037/a0018966>
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4), 525–543.

- Meredith, W., & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika*, 57(2), 289–311. <https://doi.org/10.1007/BF02294510>
- Merkle, E. C., & Rosseel, Y. (2018). blavaan: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software*, 85(4), 1–30. <https://doi.org/10.18637/jss.v085.i04>
- Millsap, R. E. (2007). Invariance in measurement and prediction revisited. *Psychometrika*, 72(4), 461–473. <http://doi.org/10.1007/s11336-007-9039-7>
- Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, 17(4), 297–334. <https://doi.org/10.1177/014662169301700401>
- 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>
- Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. *Journal of Applied Psychology*, 96(5), 966–980. <https://doi.org/10.1037/a0022955>
- Nye, C. D., Bradburn, J., Olenick, J., Bialko, C., & Drasgow, F. (2019). How big are my effects? Examining the magnitude of effect sizes in studies of measurement equivalence. *Organizational Research Methods*, 22(3), 678–709. <https://doi.org/10.1177/1094428118761122>
- Palardy, G. J. (2019). School peer non-academic skills and academic performance in high school. *Frontiers in Education (Lausanne)*, 4. <https://doi.org/10.3389/feduc.2019.00057>

- 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>
- R Core Team (2019). R: A language and environment for statistical computing. R foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.
- Radloff, L. S. (1977). The CES-D scale: A self report depression scale for research in the general population. *Applied Psychological Measurements, 1*, 385-401.
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. <http://www.jstatsoft.org/v48/i02/>.
- Schaefer, V. A. (2009). *I have a dream: Rural adolescents' educational plans and mathematics achievement*. [Unpublished doctoral dissertation]. University of North Carolina at Chapel Hill.
- Sheikh, S. (2014). Cultural variations in shame's responses: A dynamic perspective. *Personality and Social Psychology Review, 18*(4), 387–403. <https://doi.org/10.1177/1088868314540810>
- Shi, D., Song, H., Distefano, C., Maydeu-Olivares, A., Mcdaniel, H., & 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>

- Song, X. -Y., & Lee, S. -Y. (2012). A tutorial on the Bayesian approach for analyzing structural equation models. *Journal of Mathematical Psychology*, 56(3), 135–148.
<https://doi.org/10.1016/j.jmp.2012.02.001>
- Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item (functioning and differential) test functioning on selection decisions: When are statistically significant effects practically important? *Journal of Applied Psychology*, 89(3), 497-508. <http://dx.doi.org/10.1037/0021-9010.89.3.497>
- Steiger, J. H., Shapiro, A., & Browne, M. W. (1985). On the multivariate asymptotic distribution of sequential Chi-square statistics. *Psychometrika*, 50(3), 253-263.
<http://dx.doi.org/10.1007/BF02294104>
- Svetina, D., Rutkowski, L., & Rutkowski, D. (2020). Multiple-Group invariance with categorical outcomes using updated guidelines: An illustration using Mplus and the lavaan/semTools packages. *Structural Equation Modeling*, 27(1), 111–130.
<https://doi.org/10.1080/10705511.2019.1602776>
- Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, 52(3), 393-408.
- U.S. Department of Education, National Center for Education Statistics. (2004). *Education longitudinal study of 2002: Base year data file user's manual*, by Steven J. Ingels, Daniel J. Pratt, James E. Rogers, Peter H. Siegel, and Ellen S. Stutts. Washington, DC.
- Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P. (2020). Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC. *Bayesian Analysis*. <https://doi.org/10.1214/20-BA1221>

- Verhagen, A. J., & Fox, J. P. (2013). Bayesian tests of measurement invariance. *British Journal of Mathematical and Statistical Psychology*, 66(3), 383–401.
<https://doi.org/10.1111/j.2044-8317.2012.02059.x>
- Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, 54(6), 1063–1070. <https://doi.org/10.1037/0022-3514.54.6.1063>
- 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>
- Yang, L., Jia, C. -X., & Qin, P. (2015). Reliability and validity of the Center for Epidemiologic Studies Depression Scale (CES-D) among suicide attempters and comparison residents in rural China. *BMC Psychiatry*, 15(1), 76–76. <https://doi.org/10.1186/s12888-015-0458-1>
- Yuan, K. H., & Chan, W. (2016). Measurement invariance via multigroup SEM: Issues and solutions with chi-square-difference tests. *Psychological Methods*, 21(3), 405–426.
<https://doi.org/10.1037/met0000080>.

Table and Figures

Table 1

Self-Efficacy (Math-Specific) Scale Items Summary and Correlation Matrix (Spearman) by Gender

	Item Description	Female		Male		Correlation Matrix				
		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	BYS89A	BYS89B	BYS89L	BYS89R	BYS89U
BYS89A	Can do excellent job on math tests	2.43	0.92	2.68	0.92	--	.74	.68	.70	.69
BYS89B	Can understand difficult math texts	2.21	0.91	2.54	0.93	.77	--	.74	.68	.68
BYS89L	Can understand difficult math class	2.34	0.95	2.61	0.95	.70	.74	--	.75	.77
BYS89R	Can do excellent job on math assignments	2.55	0.95	2.72	0.93	.71	.68	.77	--	.81
BYS89U	Can master math class skills	2.58	0.94	2.76	0.92	.69	.69	.75	.80	--

Note. This correlation table was computed using listwise deletion, $N=10,443$. This differs from the sample size used ($N=11,663$) in *lavaan* and *blavaan* since full information maximum likelihood and Bayes estimators were used in fitting the MG-CFA models. The correlations of male (female) are shown in the lower (upper) triangle.

Table 2

Fit Statistics of Various Invariance Models

	χ^2	df	CFI	ΔCFI	RMSEA	$\Delta RMSEA$
Configural Invariance	119.862	6	0.997		0.057	
Weak Invariance	136.377	10	0.997	0.000	0.047	-0.010
Partial Weak Invariance (freed λ_5 and all τ s and θ s)	124.963	9	0.997	0.000	0.047	-0.010
Partial Strong Invariance (freed λ_5 , τ_5 , and all θ s)	310.578	12	0.993	-0.004	0.065	0.018
Partial Strong Invariance (freed λ_5 , τ_2 , τ_4 , τ_5 , and all θ s)	125.906	10	0.997	0.000	0.045	-0.002
Partial Strict Invariance (freed λ_5 , τ_2 , τ_4 , τ_5 , θ_4 , θ_5)	128.181	13	0.997	0.000	0.039	-0.006
Partial Strict Invariance (freed λ_5 , τ_2 , τ_4 , τ_5 , θ_4 , θ_5 , θ_{12})	129.045	14	0.997	0.000	0.038	-0.007

Note. The changes in fit statistics (ΔCFI and $\Delta RMSEA$) are computed between configural and partial weak invariance models, partial weak and partial strong invariance models, partial strong and partial strict invariance models.

Table 3

*Parameter Estimates for Math-specific Self-efficacy Scale Analyzed across Female and Male**Under the Partial Strict Invariance Model*

Parameter	Female		Male	
	Estimate	Posterior SD.	Estimate	Posterior SD.
<u>Factor means</u>	0		0.323	0.021
<u>Factor variances</u>	1		1.020	0.030
<u>Factor loadings</u>				
BYS89A	0.727	0.009	--	--
BYS89B	0.707	0.009	--	--
BYS89L	0.816	0.009	--	--
BYS89R	0.840	0.009	--	--
BYS89U	0.849	0.010	0.814	0.011
<u>Intercepts</u>				
BYS89A	2.430	0.011	--	--
BYS89B	2.207	0.012	2.297	0.012
BYS89L	2.337	0.012	--	--
BYS89R	2.543	0.012	2.440	0.014
BYS89U	2.567	0.012	2.479	0.014
<u>Unique factor variances</u>				
BYS89A	0.316	0.005	--	--
BYS89B	0.342	0.005	--	--
BYS89L	0.242	0.004	--	--
BYS89R	0.181	0.005	0.161	0.005
BYS89U	0.170	0.005	0.191	0.005

Note. -- represents invariant parameters, which means the cells have the same value as the other group.

Figure 1

Trace plot of the standardized factor loadings for test items

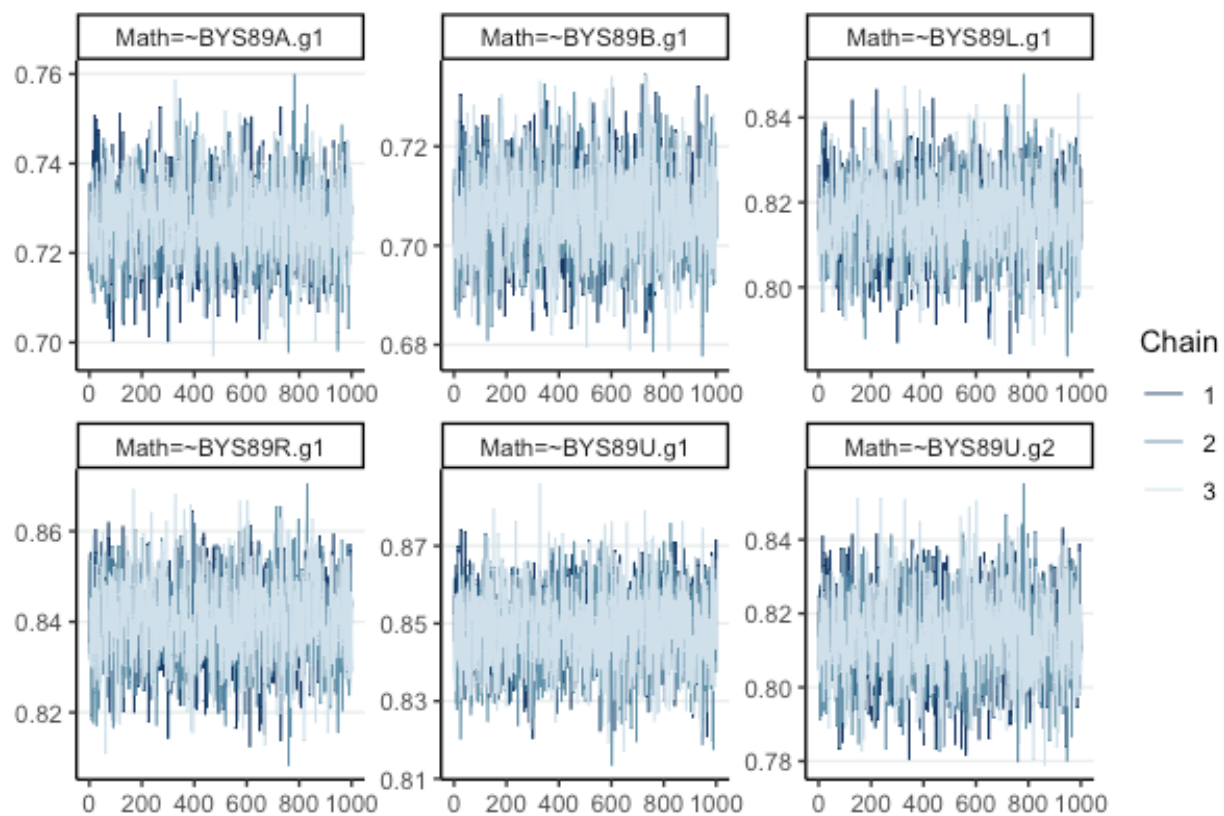
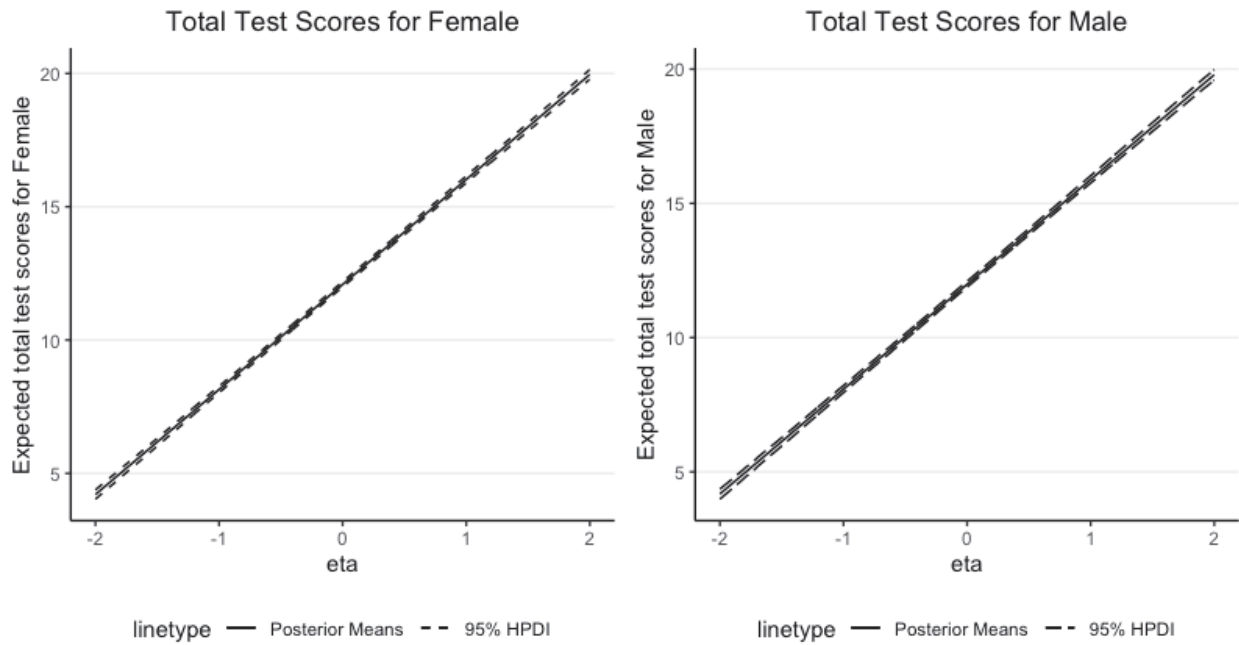
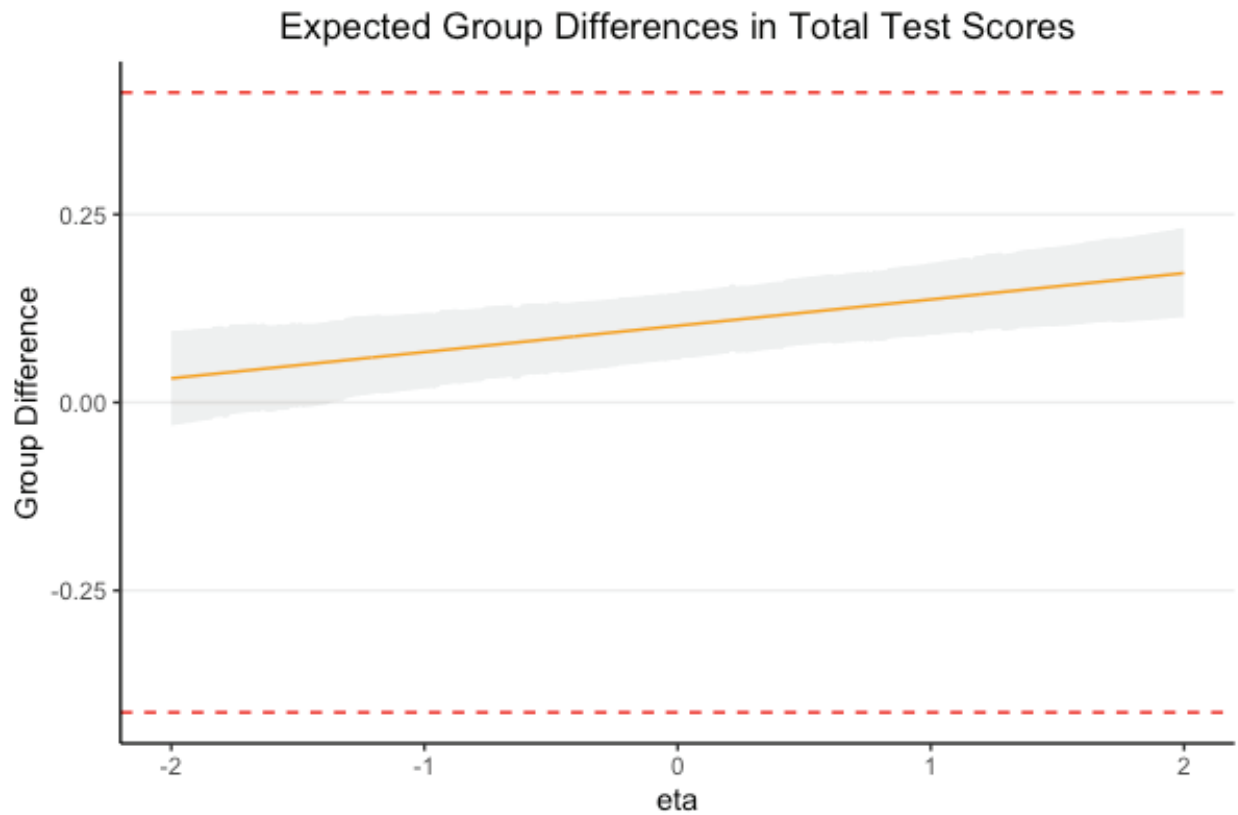


Figure 2

Expected Total Test Scores for Male and Female

Note. The black solid lines show the posterior means for female (left) and male (right), and the intervals between two dashed lines represent the 95% HPDI. The posterior means and 95% HPDI for female and male are very similar.

Figure 3

Expected Group Differences in Total Test Scores

Note. The orange (dark grey) solid line represents the posterior mean of the expected difference in total scores between female and male (female score minus male score) in math-specific self-efficacy scale. The shaded area represents the 95% HPDI, and the interval between two red (grey) dashed lines is the ROME $[-0.412, 0.412]$.

Appendix

We tested the one-factor model of math self-efficacy described in Schaefer (2009) using R (R core team, 2019), and the *lavaan* package (v0.6-5; Rosseel, 2012). This model showed a good fit: $\chi^2(df = 3, N = 11663) = 113.251, p < .001$, RMSEA = .056, 90% CI [0.048, 0.065], CFI = .998. We evaluated configural invariance across gender and found it had an acceptable fit: RMSEA = .057 (90% CI [.048, .066]), CFI = .997. However, this configural model was rejected by a test of exact fit: $\chi^2(df = 6) = 119.862, p < .001$. We then fitted a weak invariance model which constrains all factor loadings to be the same across each group to data. The Chi-square difference test was significant: $\Delta\chi^2(df = 4) = 16.515, p < .05$. Sequential likelihood ratio tests, similar to the sequential specification search proposed by Yoon & Millsap (2007) based on modification indices, were used to identify noninvariant items. Specifically, we relaxed invariance constraints of items based on a series of likelihood ratio tests until there is no significant chi-square difference. Freeing factor loading for each item sequentially suggests item 5 (“I’m certain I can master the skills being taught in my math class.”) might have different loadings for males and females.⁶ A partial weak invariance model with freely estimated factor loading for item 5 showed a good fit, with $\chi^2(df = 9) = 124.963, p < .05$, RMSEA = .047, 90% CI [0.040, 0.054], CFI = .997.

The next model tested strong invariance by further fixing intercepts for all items across each group. Again, the chi-square difference was statistically significant, suggesting some items may have noninvariant intercepts: $\Delta\chi^2(df = 4) = 207.24, p < .001$. Releasing the intercept for each item one by one suggested item 5 (“I’m certain I can master the skills being taught in my

math class”), item 4 (“I’m confident I can do an excellent job on my math assignments.”) and item 2 (“I’m certain I can understand the most difficult material presented in my math texts.”) were not measurement invariant in their intercepts for males and females. A partial strong invariance model with released intercepts for item 2, 4 and 5 was fitted to data and showed a better fit: $\chi^2(df = 10) = 125.906, p < .001$, RMSEA = .045, 90% CI [0.038, 0.052], CFI = .997.

We further constrained the unique factor variance for all items across each group. The Chi-square different test was statistically significant, indicating that some items were noninvariant on unique factor variances: $\Delta\chi^2(df = 5) = 16.741, p < .05$. We continued freeing items’ unique factor variances one by one and found item 4 (“I’m confident I can do an excellent job on my math assignments.”) and item 5 were not measurement invariant at this level. The unique factor covariances were further tested, suggesting a noninvariant covariance between items 1 and 2 across groups. Thus, our base CFA model was a partial strict invariant model with unequal unique factor variance on items 4 and 5 and unequal covariance between items 1 and 2. The final model showed a good fit: $\chi^2(df = 14) = 129.045, p < .001$, RMSEA = .038, 90% CI [0.032, 0.044], CFI = .997. The sequential investigation suggested that items 1 and 3 were strict invariant, items 2 and 4 were weak invariant, and item 5 was noninvariant.