LSAC RESEARCH REPORT SERIES

- **Marginalized Measurement Variance Modeling and Bayes Factor Testing**

Jean-Paul Fox
Vera Broks
University of Twente, Enschede, the Netherlands

- **Law School Admission Council**
Research Report 16-06
November 2016

A Publication of the Law School Admission Council

The Law School Admission Council (LSAC) is a nonprofit corporation that provides unique, state-of-the-art products and services to ease the admission process for law schools and their applicants worldwide. Currently, 222 law schools in the United States, Canada, and Australia are members of the Council and benefit from LSAC's services. All law schools approved by the American Bar Association are LSAC members. Canadian law schools recognized by a provincial or territorial law society or government agency are also members. Accredited law schools outside of the United States and Canada are eligible for membership at the discretion of the LSAC Board of Trustees; Melbourne Law School, the University of Melbourne is the first LSAC-member law school outside of North America. Many nonmember schools also take advantage of LSAC's services. For all users, LSAC strives to provide the highest quality of products, services, and customer service.

Founded in 1947, the Council is best known for administering the Law School Admission Test (LSAT®), with about 100,000 tests administered annually at testing centers worldwide. LSAC also processes academic credentials for an average of 60,000 law school applicants annually, provides essential software and information for admission offices and applicants, conducts educational conferences for law school professionals and prelaw advisors, sponsors and publishes research, funds diversity and other outreach grant programs, and publishes LSAT preparation books and law school guides, among many other services. LSAC electronic applications account for nearly all applications to ABA-approved law schools.

Table of Contents

Executive Summary

Among the assumptions that should be met when applying an item response theory (IRT) model to the analysis of test data is measurement invariance. Measurement invariance requires that, after controlling for a test taker's proficiency, group membership have no effect on the probability that that test taker will answer a test question correctly. Groups may be defined on the basis of many factors, including gender, race/ethnicity, and citizenship.

This research study proposes and evaluates a new method for detecting violations of the measurement invariance assumption. The method is evaluated through both data simulation and application to actual responses to an international survey. The results obtained by the proposed method are also compared to those obtained using the Mantel–Haenszel statistic, the industry standard for group membership comparisons. Promising results are reported, and plans for extended research are discussed.

Introduction

When administering a test to different groups, it is important to be able to compare the test results across members of those groups. In order to make meaningful comparisons between groups, the latent variable θ (i.e., ability) must be measured on a common scale. To accomplish a common scale analysis, the possible violation of the assumption of measurement invariance should be taken into account, as described by Thissen, Steinberg, and Gerrard (1986) and Fox (2010, Chapter 7). In item response theory (IRT), measurement invariance is present when the conditional probability of answering an item correctly does not depend on group information (Thissen et al., 1986).

In current Bayesian methods, random item effects are used to detect measurement variance. More specifically, deviations from the overall mean are specified for each group-specific item parameter, as described in Fox (2010, Chapter 7) and Kelcey, McGinn, and Hill (2014). The variance between groups with respect to these deviations is evaluated in order to detect measurement variance: The larger the variance between groups, the higher the degree of measurement variance. These current methods are based on a conditional IRT modeling approach, where inferences are made regarding the latent variable conditional on the estimates of the group-specific item parameters (Fox, 2010, Chapter 7). Verhagen and Fox (2013) showed that Bayesian methods can be used concurrently to test multiple invariance hypotheses for groups randomly sampled from a population. They found that a Bayes factor test had good power and low Type I error rates for different sample-size conditions to detect measurement variance. For a fixed (nonrandomly sampled), smaller number of groups, Verhagen, Levy, Millsap, and Fox (2015) proposed another Bayes factor test, which was able to directly evaluate item difficulty parameter differences among the selected groups. Moreover, van de Schoot et al. (2013) demonstrated that approximate measurement invariance can be evaluated by using a prior to determine acceptable differences between groups.

These current approaches have several limitations. First, the variance between group-specific item parameters is explicitly modeled even though the object of these methods is to test whether this variance is present, which would indicate that the measurement invariance assuption is violated (Fox, 2010, Chapter 7). That is, the prior for the variance parameter reflects an assumption of measurement variance. Second, the model representing measurement invariance is not nested within the model representing measurement variance. Measurement invariance is represented by a variance of zero, which is a boundary value on the parameter space (Fox, Sinharay, & Mulder, 2016). This complicates statistical test procedures and requires approximate methods such as an encompassing prior approach (Klugkist & Hoijtink, 2007). Third, the latent variable θ is estimated using potentially biased item difficulty and population parameter estimates. Fourth, the above-mentioned approaches are applicable either to a fixed (nonrandomly selected) number of groups or to randomly selected groups, but none of the approaches is applicable to both situations.

To overcome these limitations, a new method based on a marginalized item response model is proposed. Instead of conditioning on group-specific item parameters, a common item difficulty parameter is modeled, which applies to all groups. As a result, the possible error with respect to this item difficulty parameter is included in the residuals for each group. It is proposed that in order to detect measurement variance, the correlation of within-group residuals should be evaluated. Hence, the additional correlation between observations caused by violations of measurement invariance is addressed in the marginalized item response model. Additionally, since residual correlations between response probabilities are evaluated, the complex identification assumptions associated with the random item effects model (De Jong, Steenkamp, & Fox, 2007; Verhagen & Fox, 2013) can be avoided. A further benefit of the proposed method is that it can be applied to both randomly and nonrandomly selected groups.

In the following sections, the marginalized item response model is explained and then evaluated in a simulation study with a fixed number of groups. The fractional Bayes factor is used to objectively compare competing hypotheses to accommodate an improper prior for the implied degree of measurement variance. The functioning of the fractional Bayes factor will be compared to that of the posterior predictive check based on the Mantel–Haenszel chi-square statistic χ^2_{MH} to evaluate measurement invariance assumptions, since the latter is a commonly used tool to detect measurement variance (Holland & Wainer, 1993). The fractional Bayes factor has many advantages over the χ^2_{MH} statistic and thus performs better. The proposed method is then extended in order to be applicable to a larger number of randomly selected groups, for which parameter recovery is also evaluated in a simulation study. Finally, the method is applied to empirical data using data from the European Social Survey (ESS).

Marginalized Item Response Model

In a conditional modeling approach, group-specific item parameters are modeled. For instance, the random item effects model is a conditional model in which a normal distribution is assumed for the group-specific item parameters. This model has been used by Verhagen and Fox (2013) and De Jong et al. (2007) to detect violations of measurement invariance. In this report, it is shown that a marginal model can be derived from this random item effects model such that group-specific item parameters are no longer modeled. The one-parameter multilevel IRT model will be used for illustration, as described by Bock and Zimowski (1997) and Azevado, Andrade, and Fox (2012), and the probability of answering an item correctly is given by

$$P(Y_{ijk} = 1 \mid \theta_{ij}, b_k) = \frac{\exp(\theta_{ij} - b_k)}{1 + \exp(\theta_{ij} - b_k)}, \tag{1}$$

where θ_{ij} is the underlying ability of person i in group j, and b_k is the difficulty of item k. Parameter b_k reflects the required value of the underlying ability θ in order for the test taker to have an expected probability of .5 of answering the item correctly.

The Random Item Effects Model

Assume for the moment continuous responses to items, symbolized by Z_{ijk}. In a random item effects model, this latent response variable is modeled as follows:

$$Z_{ijk} = \theta_{ij} - b_{jk} + \varepsilon_{ijk,} \varepsilon_{ijk} \sim N(0,1), \tag{2}$$

where

$$b_{jk} = b_k + \varepsilon_{jk}, \varepsilon_{jk} \sim N(0, \tau_k). \tag{3}$$

In Equation (2), the random item effects model is shown, where the latent response variable Z_{ijk} is independently and identically distributed given the group-specific item difficulty parameter b_{jk} and person parameter θ_{ij}. As illustrated in Equation (3), the random item effects parameter is assumed to be normally distributed with the mean equal to the invarant item difficulty parameter b_k and variance τ_k. The variance parameter τ_k stands for the between-group variance with respect to the random item difficulty parameter, and it represents the degree of measurement variance.

The Marginal Random Item Effects Model

The random item effects model can be marginalized by integrating out the group-specific item parameters. This can be done by plugging Equation (3) into Equation (2). It follows that

$$\begin{aligned} Z_{ijk} &= \theta_{ij} - b_{jk} + \varepsilon_{ijk} \\ &= \theta_{ij} - b_k + \varepsilon_{ijk} + \varepsilon_{jk} , \\ &= \theta_{ij} - b_k + E_{ijk} , \end{aligned} \tag{4}$$

where the errors for item k ($\mathbf{E}_k$) are assumed to have a multivariate normal distribution with a mean of zero and covariance matrix $\mathbf{\Sigma}_k$.

In this marginal model, the latent response variable no longer depends on the group-specific item parameter b_{jk}, and one item difficulty parameter b_k applies to all groups. As a result, the degree of measurement variance is included in the error term. Note that in this marginal model, conditional independence no longer applies due to the fact that group-specific item parameters are not specified. In the marginalized item response model as described here, Z_{ijk} has a multivariate normal distribution. The presence of measurement variance is absorbed into the covariance structure of the error term.

To explain the covariance structure of the marginal model in more detail, let $\varepsilon_{ijk} \sim N\left(0, \sigma^2_{\varepsilon k}\right)$, which means that the measurement error variance in Equation (2) is $\sigma^2_{\varepsilon k}$. Then $\mathbf{\Sigma}_k$ can be specified. In the first case, $i = i'$, which automatically implies that $j = j'$. This reflects the covariance of two responses of person i in group j. In the second case, $i \neq i'$ but $j = j'$. That is, different persons i and i' are in the same group j. The third case consists of the covariance of different persons i and i' in different groups j and j'. From Equation (5) it can be concluded that the (co)variances in these three different cases are equal to $\tau_k + \sigma^2_{\varepsilon k}$, τ_k, and 0, respectively:

$$\begin{aligned} \mathbf{\Sigma}_k &= \operatorname{cov}(\varepsilon_{jk} + \varepsilon_{ijk}, \varepsilon_{j'k} + \varepsilon_{i'j'k}) \\ &= \operatorname{cov}(\varepsilon_{jk}, \varepsilon_{j'k}) + \operatorname{cov}(\varepsilon_{ijk}, \varepsilon_{i'j'k}) \\ &= \begin{cases} \operatorname{var}(\varepsilon_{jk}) + \operatorname{var}(\varepsilon_{ijk}) = \tau_k + \sigma^2_{\varepsilon k} & \text{if } i = i', j = j' \\ \operatorname{var}(\varepsilon_{jk}) = \tau_k & \text{if } i \neq i', j = j' \\ 0 & \text{if } j \neq j'. \end{cases} \end{aligned} \tag{5}$$

In this marginal model, there is only one item difficulty parameter present, which applies to all the groups, instead of there being an item difficulty parameter for each group separately. The possible error due to measurement variance is no longer explicitly modeled in b_k but included in the covariance structure of the error distribution. In $\mathbf{\Sigma}_k$ the presence of measurement variance is captured by the covariance of different observations within a group, specified by the second case in Equation (5). It is proposed that in order to test whether measurement variance is present (and to what degree), one should evaluate τ_k.

When the groups are randomly selected from a population, the covariance structure for the responses to item *k* in group *j* is given by

$$\mathbf{\Sigma}_{jk} = \sigma^2{}_{\varepsilon k}\mathbf{I}_m + \tau_k\mathbf{J}_m, \tag{6}$$

where $\mathbf{I}_m$ is the identity matrix and $\mathbf{J}_m$ a matrix of ones; m stands for the number of observations in each group, and equal group sizes (balanced design) are assumed. In the covariance structure of Equation (6), parameters τ_k on the off-diagonal positions represent the implied covariance between latent responses due to the clustering of responses in groups. Parameters τ_k on the diagonal positions contribute to the variance in item difficulty across groups. For randomly selected groups, the random item effect parameter is used to model the clustering of responses in groups as well as the variability in item functioning across groups. The groups are sampled from a population, and the random item effects variance represents the variance in item functioning in the population of groups. For all items *k*, when binary response data are observed, the variance parameter $\sigma^2_{\varepsilon k}$ will be fixed to one to identify the scale.

For a fixed number of groups, the variability across groups does not apply, since the groups are not sampled from a population. Then, the total variance is the sum of the measurement error variances and the covariances. Another parameterization is used to avoid the situation where the covariance parameters can model any variability between groups, as in the situation for randomly selected groups.

To accomplish this, a parameterization of the covariance matrix is presented, where the covariance parameters can only modify the covariance between parameters. The total variance of the scale is determined by the sum of the variance components $\sigma^2_{\varepsilon k}$ and the covariance components τ_k. Therefore, to restrict the additional contribution of the covariance components to the total variance, the variance parameter $\sigma^2_{\varepsilon k} = 1 - \tau_k$. In that case the total variance is always equal to one, and the covariance components are not allowed to increase the total variance. Note that in the parameterization presented in Equation (6), the covariance parameters can modify the covariance between response observations as well as the total amount of variance in response observations.

For a fixed number of groups, parameter τ_k should only model the within-group covariance and not any variance in item functioning across groups. For a fixed number of groups, the covariance structure is adapted and $\sigma^2_{\varepsilon k}$ is restricted to be equal to $1-\tau_k$, which reduces the covariance matrix of the error terms for each group *j* and item *k* to

$$\mathbf{\Sigma}_{jk} = (1-\tau_k)\mathbf{I}_m + \tau_k \mathbf{J}_m . \tag{7}$$

It follows that the values on the diagonal are equal to 1 and the off-diagonal values are equal to τ_k. In this covariance structure $\mathbf{\Sigma}_{jk}$, the τ_k is a correlation parameter, since the diagonal consists of ones.

Priors and the MCMC Algorithm

In order to estimate the degree of measurement variance under the marginalized item response model, a Markov chain Monte Carlo (MCMC) algorithm is presented in which samples are iteratively drawn from conditional distributions. This process is illustrated in Figure 1; a detailed outline of this MCMC algorithm can be found in Appendix A. The priors and sampling steps specifically defined for the considered marginalized item response model are discussed, with the remaining steps provided in Appendix A.

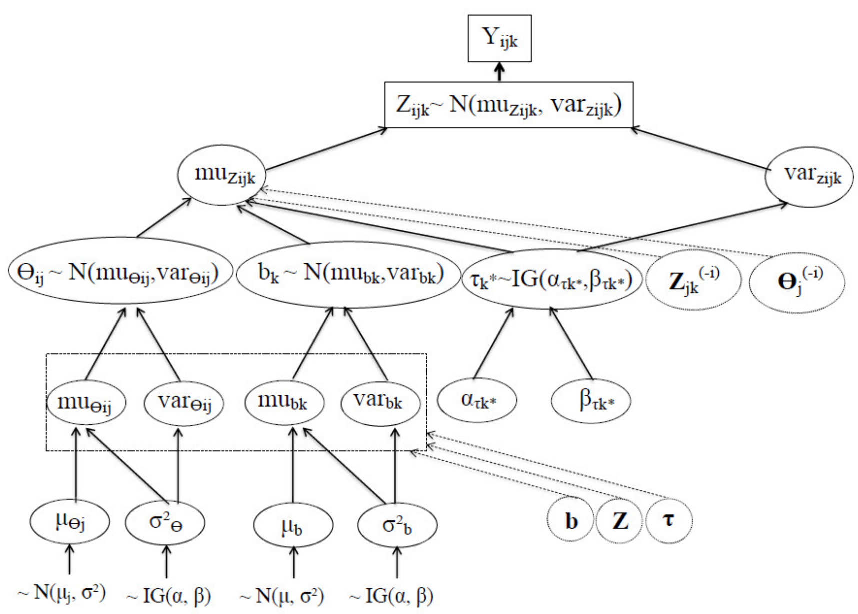

FIGURE 1. *Illustration of the MCMC algorithm to sample* Z_{ijk}, θ_{ij}, b_k, and τ_k

Sample Latent Response Data

At the top of Figure 1, the observed data Y_{ijk} represents the dichotomous outcome for person i, from group j, for item k. This outcome is equal to zero when the answer is incorrect and equal to one when the answer is correct. When Y_{ijk} = 0, the latent response variable Z_{ijk} is modeled to be less than zero. When Y_{ijk} = 1, Z_{ijk} is modeled to be greater than zero. Hence, in order to accomodate this binary response data, Z_{ijk} is sampled from a truncated normal distribution with a mean and variance denoted by $mu_{Z_{ijk}}$ and $var_{Z_{ijk}}$, respectively:

$$Z_{ijk} \mid \theta_{ij}, b_k, \tau_k, \mathbf{Z}_{jk}^{(-i)}, \boldsymbol{\theta}_{jk}^{(-i)} \sim N(mu_{z_{ijk}}, var_{z_{ijk}}), \qquad (8)$$

where $Z_{ijk} > 0$ ($Z_{ijk} \leq 0$) when $Y_{ijk} = 1$ ($Y_{ijk} = 0$). As shown in Equation (8), the sampling of Z_{ijk} depends on θ_{ij}, b_k, and τ_k, and the values of $\mathbf{Z}_{jk}^{(-i)}$ and $\boldsymbol{\theta}_j^{(-i)}$, which indicates the values on these parameters for the other members of the same group j.

Sample Person Parameters

Here, a multilevel normal population distributuion is assumed for the person parameters, with a group-specific intercept μ_{θ_j} and variance σ_θ^2. The hyperparameters μ_{θ_j} and σ_θ^2 have hyperpriors, which are given by

$$\mu_{\theta_j} \sim N(\mu_j, \sigma^2)$$

and

$$\sigma_\theta^2 \sim IG(\alpha, \beta),$$

respectively. Subsequently, as illustrated by Figure 1, the person parameter θ_{ij} is sampled from a normal distribution with mean $mu_{\theta_{ij}}$ and variance $var_{\theta_{ij}}$:

$$\theta_{ij} \mid \mathbf{Z}_{ij}, \mathbf{b}, \boldsymbol{\tau}, \mu_{\theta_j}, \sigma_\theta^2 \sim N(mu_{\theta_{ij}}, var_{\theta_{ij}}). \qquad (9)$$

Sample Item Parameters

The prior parameters μ_b and σ_b^2 have a normal-inverse-gamma distribution, which is specified as follows:

$$\mu_b \sim N(0, \sigma_b^2 / n_0)$$

$$\sigma_b^2 \sim IG(\alpha_b, \beta_b),$$

where $n_0 \geq 0$ determines the weight of the prior specification of μ_b. Then, item difficulty parameter b_k is sampled from a normal distribution with mean mu_{b_k} and variance var_{b_k}:

$$b_{k,} \mid \mathbf{Z}_k, \tau_k, \mu_b, \sigma_b^2 \sim N(mu_{b_k}, var_{b_k}). \tag{10}$$

Sample Covariance Parameters

Given the covariance structure in Equation (6), priors are specfied for the variance parameter $\sigma_{\varepsilon k}^2$ and covariance parameter τ_k. Fox et al. (2016) showed that τ_k should be greater than $\sigma_{\varepsilon k}^2 / m$ to have a positive-definite covariance matrix. Therefore, a uniform improper prior for τ_k and an inverse-gamma prior for the variance parameter $\sigma_{\varepsilon k}^2$ are specified as

$$p(\tau_k \mid \sigma_{\varepsilon k}^2) \propto (\sigma_{\varepsilon k}^2 / m + \tau_k)^{-1} \tag{11}$$

$$p(\sigma_{\varepsilon k}^2) \sim IG(\alpha_\sigma, \beta_\sigma).$$

It follows then that $\sigma_{\varepsilon k}^2$, given that $\mathbf{Z}_k, \boldsymbol{\theta}, b_k$ can be sampled from an inverse-gamma distribution. Fox et al. (2016) showed that $\tau_k^* = \sigma_{\varepsilon k}^2 / m + \tau_k$ can be sampled from an inverse gamma distribution to obtain samples of τ_k.

Given the covariance structure in Equation (7), the prior for τ_k is given by

$$p(\tau_k) \propto ((m-1)^{-1} + \tau_k)^{-1}, \tag{12}$$

where $\tau_k \geq -1/(m-1)$ in order to have a positive-definite covariance matrix. Furthermore, $\tau_k \leq 1/2$ in order to restrict the level of measurement variance to be less than the error variance. Let $\sigma^2 = 1-\tau_k$; then the prior for $\sigma^2_{\varepsilon k}$ is given by

$$p(\sigma^2_{\varepsilon k}) \propto (\sigma^2_{\varepsilon k})^{-1}, \tag{13}$$

where $1/2 \leq \sigma^2{}_{\varepsilon k} \leq 1+1/(m-1)$. As shown in Fox et al. (2016), $\tau_k + (m-1)^{-1}$ has an inverse-gamma distribution and depends on $\mathbf{Z}_k, \boldsymbol{\theta}, b_k$. The variance parameter $\sigma^2_{\varepsilon k}$ can also be sampled from an inverse-gamma distribution.

When sampling the $\sigma^2_{\varepsilon k} = 1-\tau_k$ and τ_k in different steps, we found that the statistical inferences on the sampled values are complicated, since the sum of both components is restricted to one. A more efficient inference about τ_k can be made when the posterior information about $\sigma^2_{\varepsilon k} = 1-\tau_k$ is included. Therefore, the intraclass-correlation coefficient is considered, which is given by $\tau_k/(\tau_k + \sigma^2_{\varepsilon k})$. For the covariance structure defined in Equation (7), it is equal to the correlation coefficient τ_k. The intraclass correlation is not scale dependent, which makes it possible to sequentially sample a value for τ_k and $\sigma^2_{\varepsilon k}$, without the restriction that the sampled values sum to one. As a result, each computed intraclass correlation given the sampled parameter values is a sampled value of parameter τ_k, and given the sampled values, posterior inferences can be made about τ_k.

Simulation Study for Fixed Groups

In this simulation study, parameter recovery of the marginalized item response model was evaluated for the situation of a fixed number of groups, that is, for the covariance structure given by Equation (7). The first goal of this simulation study was to test whether the marginalized item response model is able to accurately estimate the degree of measurement variance. The second goal of this simulation study was to evaluate the use of the (fractional) Bayes factor to decide whether or not the degree of measurement variance in an item is equal to zero.

The fractional Bayes factor was used to accommodate for the improper prior for the measurement variance parameter τ_k (see Equations [11] and [12]). This improper prior assumes a uniform distribution for the possible degrees of measurement variance, which makes it possible to objectively evaluate the measurement invariance assumption. The fractional Bayes factor approach has several important advantages. First, it is able to test for measurement variance in all of the items simultaneously and does not require a sequential test procedure in which items are tested one by one. Second, anchor items are not needed, and full measurement invariance can also be tested using the same procedure. Third, it takes into account both the null hypothesis (H_0), which states that measurement invariance holds, as well as the alternative hypothesis (H_1), which states that measurement invariance does not hold. The posterior predictive *p*-value based on the Mantel–Haenszel χ^2_{MH} statistic (ppp χ^2_{MH}) does not have these advantages, but the functioning of the ppp χ^2_{MH} is compared to the functioning of the fractional Bayes factor.

Method

In order to evaluate the marginalized item response model with respect to estimating and detecting measurement variance, a simulation study was conducted using our own program developed in R (R Core Team, 2014). In this simulation study, binary response data were simulated for 10 items, with 1,000 persons assigned to one of two groups. The degree of measurement variance τ_k was increased across items. The lower bound of $\tau_k = -1/m$, where $m = 500$ represents the number of persons per group. For item 1, the level of measurement variance equaled this lower bound. The simulation study consisted of 50 data replications, which provided stable results; the mean results across replications are reported.

The MCMC algorithm (see section titled Priors and the MCMC Algorithm) was used to estimate the degree of measurement variance in each item. The number of MCMC iterations was set to 5,000 with a burn-in of 1,000. The convergence and autocorrelation plots, created using the R package coda (Plummer, Best, Cowles, & Vines, 2006), showed no irregularities. The functioning of the fractional Bayes factor was compared to the functioning of the ppp χ^2_{MH} statistic for the detection of measurement variance. The two approaches are discussed in more detail below.

Fractional Bayes Factor

The Bayes factor is used to evaluate the support of the data for H_0 compared to the support of the data for H_1 (Kass & Raftery, 1995; Raftery, 1995). The Bayes factor is computed as the probability of the data given the null hypothesis divided by the probability of the data given the alternative hypothesis:

$$\mathrm{BF}_{01} = \frac{p\left(\mathbf{y} \middle| \mathrm{H}_0\right)}{p\left(\mathbf{y} \middle| \mathrm{H}_1\right)}. \tag{14}$$

The probabilities in Equation (14) are referred to as the marginal distribution of the data, given the model under the concerning hypothesis. The marginal likelihood of the data given hypothesis H_i can be specified as follows (Raftery, 1995):

$$p\left(\mathbf{y} \mid \mathrm{H_i}\right)=\int p\left(\mathbf{y} \mid \omega_i, \mathrm{H_i}\right) p\left(\omega_i \mid \mathrm{H_i}\right) d\omega_i . \tag{15}$$

Here, ω_i stands for the parameters in the model under hypothesis H_i. To make objective decisions about the level of measurement variance, improper priors are specified for the covariance parameter τ_k, as specified in Equations (11) and (12), leading to an expression of the marginal distribution of the data up to an unknown constant. The corresponding outcome of the Bayes factor cannot be interpreted, since it depends on an unknown constant. To take the improper prior into account, the fractional Bayes factor is evaluated (O'Hagan, 1995).

In Appendix B, a more detailed description is given of the fractional Bayes factor for the marginalized item response model to evaluate data evidence in favor of the null hypothesis compared to the alternative hypothesis. Following Fox et al. (2016), analytical expressions of fractional Bayes factors are given to evaluate measurement invariance hypotheses. For randomly selected groups and for fixed groups, fractional Bayes factors are computed to evaluate H_0: τ =0 and H_1: $\tau \neq$0, which represents the null hypothesis that the item is measurement invariant and the alternative hypothesis that the item is not measurement invariant, respectively. For an item that cannot be characterized as measurement invariant, it is possible that (a) the item is measurement variant and shows differential item function across groups (i.e., item responses are group-specific positively correlated), or (b) the item does not contribute to the measurement scale (item responses are group-specific negatively correlated). The fractional Bayes factor, denoted as FBF_{01}, evaluates the evidence in favor of measurement invariance (H_0: τ =0) against the hypothesis that there is no measurement invariance (H_1: $\tau \neq$0). Another fractional Bayes factor is considered and referred to as FBF_{02}, which evaluates the evidence in favor of measurement invariance (H_0: τ =0) against the alternative hypothesis that there is measurement variance (H_2: τ >0), such that a negative τ_k is not supported by either the null hypothesis or the alternative hypothesis. In this case, the data is used to evaluate the evidence in favor of measurement invariance or in favor of measurement variance.

The interpretation of the resulting fractional Bayes factor remains the same as the interpretation of the Bayes factor. The guidelines for this interpretation are followed as stated by Kass and Raftery (1995). For the Bayes factor, specified in Equation (14), where the marginal distribution of the data under the null hypothesis is given in the numerator, a (fractional) Bayes factor between 0 and .333 indicates that there is three times more evidence against the null hypothesis. A (fractional) Bayes factor greater than .333 indicates that there is no substantial evidence against the null hypothesis and that measurement invariance is present.

Mantel–Haenszel Statistic: Posterior Predictive Check

The χ^2_{MH} statistic is a commonly used tool to detect measurement variance (Holland & Wainer, 1993). It can be computed to detect measurement variance between two groups, where one group is called the *reference group* and the other group is called the *focal group*. In order to compute the χ^2_{MH} statistic, the persons are divided over subgroups g based on their total test score. In a test with 10 items, this entails that the total number of subgroups G = 11, since a total score from 0 to 10 is possible. Subsequently, a contingency table can be created for each subgroup g (Hambleton & Rogers, 1989):

	Incorrect (0)	Correct (1)	
Reference Group	A_g	B_g	A_g+B_g
Focal Group	C_g	D_g	C_g+D_g
	A_g+C_g	B_g+D_g	N_g

FIGURE 2. *Contingency table for subgroups g*

The contingency table in Figure 2 is used to calculate the χ^2_{MH} statistic, which can be retrieved from the R package difR (Magis, Beland, & Raiche, 2015). The χ^2_{MH} statistic is computed as follows (Hambleton & Rogers, 1989; Magis et al., 2015):

$$\chi^2_{\mathrm{MH}} = \frac{\left(\left|\sum_{g=1}^{G} A_g - \sum_{g=1}^{G} E(A_g)\right| - \frac{1}{2}\right)^2}{\sum_{g=1}^{G} V(A_g)}, \tag{16}$$

where

$$E\left(A_g\right) = \frac{\left(A_g + C_g\right)\cdot\left(A_g + B_g\right)}{N_g} \tag{17}$$

and

$$V\left(A_g\right)=\frac{\left(A_g+C_g\right)\cdot\left(B_g+D_g\right)\cdot\left(A_g+B_g\right)\cdot\left(C_g+D_g\right)}{N_g^2\cdot\left(N_g-1\right)}. \qquad (18)$$

Sinharay, Johnson, and Stern (2006) showed that the χ^2_{MH} statistic is useful in assessing model fit in posterior predictive model checking. They used the χ^2_{MH} statistic in order to test for local independence, where responses to items are assumed to be independently distributed given the person parameter. The association among item pairs was investigated to detect possible violations of the local independence assumption. This relates to the assumption of measurement invariance, where responses to item k are assumed to be independently distributed given a common item difficulty parameter for the reference and focal groups. Therefore, it is to be expected that the statistic can also be used to test measurement invariance assumptions. When responses to item k are independently distributed given the item parameter and group membership of the respondents (i.e., reference or focal group), it is concluded that measurement invariance does not hold.

The χ^2_{MH} statistic will be used as a discrepancy measure in a posterior predictive check in order to evaluate the measurement invariance assumption. The computation of the χ^2_{MH} statistic is specified in Equation (16). Since the χ^2_{MH} statistic needs anchor items, all the other items of the test are chosen as anchor items. The object is to identify which items are measurement invariant. Unlike the χ^2_{MH} statistic, the fractional Bayes factor can do this for each item without needing information about the measurement invariance assumptions of the other items. Here, a conservative approach is followed where each item is tested by assuming the other items to be measurement invariant. In practice, it is usually not known which items are measurement invariant, and so an assumption needs to be made in order to test a single item.

Data are replicated under the model, where it is assumed that the degree of measurement variance $\tau = 0$. The posterior predictive p-value (ppp χ^2_{MH}) is estimated by the proportion of MCMC iterations in which the value of the χ^2_{MH} statistic for the replicated data is greater than the one for the observed data:

$$P\left(\chi^2_{\mathrm{MH}}\left(\mathbf{y}_{\mathrm{rep}_t}\right)\geq\chi^2_{\mathrm{MH}}\left(\mathbf{y}_{obs}\right)\mid\mathbf{y}_{obs}\right). \qquad (19)$$

The simulation study involves 50 data replications, and the mean of the ppp χ^2_{MH} over 50 replications is computed. The estimated ppp χ^2_{MH} represents the extremeness of the statistic for the observed data using replicated data generated under the assumption of measurement invariance. When the observed statistic value is extreme under the assumption of measurement invariance, a violation of this assumption is detected. A ppp χ^2_{MH} of .5 indicates that the measurement invariance assumption is not violated, whereas a value close to 0 indicates that it is (Sinharay et al., 2006). However, as Gelman, Meng, and Stern (1996) pointed out, the ppp χ^2_{MH} shows the degree to which there are discrepancies between the model and the observed data. They emphasize that it is more of a tool to assess the usefulness of a model than a test to determine whether or not the model is true.

An interesting note here is the apparent similarity between the Mantel–Haenszel test and the proposed Bayes factor test based on the marginalized item response model. Both methods evaluate a dependency between group membership and observed item responses. When measurement invariance holds, responses within each group (i.e., focal and reference) are assumed to be conditionally independently distributed given the common difficulty level. In that case, the randomly selected responses are a simple random sample. Subsequently, a violation of measurement invariance corresponds to a violation of the basic assumption of independence of a simple random sample. Both methods evaluate whether there is an interaction between the group and the item responses, which corresponds to evaluating the independence assumption of the simple random sample. If this assumption is violated, a cluster (or stratified) sample is obtained instead of a simple random sample, and measurement invariance does not hold.

Results

Table 1 presents the results of the simulation study. In this simulation, measurement variance increases across items. Parameter τ represents the simulated degree of measurement variance whereas τ' represents the estimated degree of measurement variance by the posterior mean. Column $\tau - \tau'$ shows the difference between the simulated measurement variance and the estimated measurement variance by the posterior mean computed under the marginalized item response model. Table 1 shows that the estimated degree of measurement variance τ' differs a maximum of .089 from the simulated degree of measurement variance τ. The smallest absolute difference between the two values is equal to .001. It appears here that when the degree of measurement variance is smaller than .075, the posterior mean, as a point estimate, tends to overestimate the degree of measurement variance. When the degree of measurement variance is greater than 0.100 it tends to underestimate the degree of measurement variance.

TABLE 1

Fixed groups: Results of the simulation study, replicated 50 times, for estimating the degree of measurement variance

				Fractional Bayes Factor				Posterior Predictive Check		
Item	τ	τ'	$\tau-\tau'$	ln(FBF$_{01}$)	FBF$_{01}$	ln(FBF$_{02}$)	FBF$_{02}$	ppp χ^2_{MH}	Range ppp χ^2_{MH}	%ppp < 0.05
1	−0.002	0.033	−0.035	−0.755	0.470	−0.158	0.853	0.282	[0.000, 0.908]	28
2	0.000	0.038	−0.038	−1.208	0.299	−0.646	0.524	0.309	[0.000, 0.911]	36
3	0.025	0.047	−0.022	−2.463	0.085	−2.009	0.134	0.251	[0.000, 0.984]	44
4	0.050	0.077	−0.027	−7.980	<0.001	−7.686	<0.001	0.129	[0.000, 0.919]	60
5	0.075	0.076	−0.001	−7.036	0.001	−6.803	0.001	0.094	[0.000, 0.717]	68
6	0.100	0.097	0.003	−13.659	<0.001	−13.415	<0.001	0.111	[0.000, 0.776]	70
7	0.125	0.095	0.030	−11.875	<0.001	−11.703	<0.001	0.070	[0.000, 0.908]	80
8	0.150	0.099	0.051	−13.128	<0.001	−12.927	<0.001	0.100	[0.000, 0.869]	76
9	0.175	0.114	0.061	−18.977	<0.001	−18.791	<0.001	0.075	[0.000, 0.833]	78
10	0.200	0.111	0.089	−18.864	<0.001	−18.711	<0.001	0.068	[0.000, 0.890]	80

FBF$_{01}$ = fractional Bayes factor, where H$_0$: τ =0 and H$_1$: $\tau \neq 0$; FBF$_{02}$ = fractional Bayes factor, where H$_0$: τ =0 and H$_2$: τ >0; ppp χ^2_{MH} = posterior predictive *p*-value based on the Mantel–Haenszel χ^2_{MH} statistic; ppp χ^2_{MH} = mean of the posterior predictive *p*-values over the 50 replications; Range ppp χ^2_{MH} = range of the found posterior predictive *p*-values over the 50 replications; %ppp < 0.05 shows the percentage of the 50 replications that resulted in a posterior predictive *p*-value based on ppp χ^2_{MH} < .05.

The posterior mean estimate of the variance parameter differs from the mode, since the posterior distribution is skewed. For a small (large) variance parameter, the posterior is skewed to the right (left) and the mean is higher (lower) than the posterior mode and over-(under) estimates the true value. For this reason, evaluating the presence of measurement variance using point estimates is not recommended. In order to test whether the estimated degree of measurement variance $\tau' = 0$ or not, a fractional Bayes factor is computed and compared to the functioning of the ppp χ^2_{MH} for model selection. To compute the fractional Bayes factor, posterior samples are used, not point estimates, because posterior samples take the skewness of the posterior into account.

In the marginalized item response model, the natural logarithm of the fractional Bayes factor is computed (see Table 1). These results can be found under columns ln(FBF_{01}) and ln(FBF_{02}), which show that the higher the degree of measurement, the more negative the natural logarithm of the fractional Bayes factor. Table 1 also shows the fractional Bayes factors (columns FBF_{01} and FBF_{02}). Though FBF_{02} performs very well for all the items, it is greater than .333 for items 1 and 2. Therefore, it can be concluded that items 1 and 2 are measurement invariant. However, FBF_{01} shows support for the alternative hypothesis (H_1) for item 2, while this actually is a measurement invariant item. So, FBF_{02} performs better than FBF_{01} in deciding whether item 2 is measurement invariant. This can be explained as follows. FBF_{02} results for items 1 and 2 show more support for the measurement invariance hypothesis (H_0) than $\mathrm{FBF}_{01,}$ since the alternative hypothesis is restricted to the measurement variance hypothesis (H_2). Therefore, support for small, negative values of τ_k do not contribute to evidence in favor of alternative hypothesis H_2 ($\tau_k > 0$), whereas those values do contribute to alternative hypothesis H_1 ($\tau_k \neq 0$). So, more power was obtained in detecting measurement invariance by restricting the alternative hypothesis to measurement variance (H_2).

Hypotheses H_0 and H_1 were equally likely for items 1 and 2, but the fractional Bayes factors were not equal to one. The alternative hypothesis H_1 also covers τ values, which are close to, but not exactly equal to, zero. The data give the most support to τ values equal to or close to zero, which makes H_1 slightly more attractive than H_0.

For items 3–10, the fractional Bayes factors indicate that measurement variance is present, and this was also simulated for these items. However, note that alternative hypothesis H_2 represents measurement variance, whereas the evidence in favor of alternative hypothesis H_1 represents all values of $\tau_k \neq 0$. For instance, for item 3, it is 11.76 (1/0.85) more likely that $\tau_3 \neq 0$, but only 7.46 (1/.134) more likely that $\tau_3 > 0$, which represents measurement variance.

The results of the ppp χ^2_{MH} can be found in the last three columns of Table 1. It's hard to draw conclusions based on these values, since there is no common cut-off score. A ppp χ^2_{MH} close to zero shows discrepancies between the model that assumes measurement invariance and the observed data. Items 1–3 appear to have a smaller degree of discrepancy; items 4–10 appear to have a larger degree of discrepancy, since these values are closer to zero. This result is not exactly in line with that for the simulated data, since measurement variance was also present in item 3, which could not be clearly concluded from the results of the ppp χ^2_{MH}. Column %ppp < 0.05 shows the percentage of the 50 replications in which ppp χ^2_{MH} values were extreme (i.e., close to zero). Here, ppp χ^2_{MH} values are interpreted as extreme when they are less than .05. This column is provided to offer more insight with respect to the distribution of the ppp χ^2_{MH}. It is not meant as a threshold value for either accepting or rejecting the model. From this column it can be concluded that for items 1–3, ppp χ^2_{MH} < .05 for less than half of the 50 replications, and for items 4–10, ppp χ^2_{MH} < .05 for more than half of the replications.

The fractional Bayes factor has considerable benefits compared to the ppp χ^2_{MH} statistic. First, the fractional Bayes factor is able to test the degree of measurement variance for all the items at once, without the need to specify anchor items. Second, it compares the probability of the data given the null hypothesis to the probability of the data given the alternative hypothesis. This entails that both hypotheses are evaluated and the degree of support for each of them is compared. Consequently, the results are easy to interpret, since they either provide a preference for one of the two models or indicate that there is no preferable model. Finally, unlike the χ^2_{MH} statistic, which is only applicable for the comparison of two groups, the fractional Bayes factor can be computed for two or more groups. As expected, the results of the fractional Bayes factor are more convincing compared to those of the χ^2_{MH} statistic. Together with the other benefits of the fractional Bayes factor, it appears that this is an improved tool for detecting the presence of measurement variance.

Simulation Study for Random Groups

In this simulation study, parameter recovery by the marginalized item response model was evaluated in the situation where groups are randomly selected from a larger population. The covariance structure defined in Equation (6) was assumed for the responses to item *k*. The corresponding marginalized item response model was tested by estimating the degree of measurement variance τ_k for every item; $\sigma^2_{\varepsilon k}$ = 1 to identify the scale. Furthermore, the fractional Bayes factor (defined in Appendix B) was used to quantify the evidence against the hypothesis that the degree of measurement variance is equal to zero.

Method

In order to evaluate the power of the fractional Bayes factor for detecting measurement variance in the situation where groups are randomly selected, a simulation study was conducted using software developed in R (R Core Team, 2014). In this study, a dataset was generated with 1,000 persons, equally divided over 20 randomly selected groups. The responses (either incorrect or correct) of these 1,000 persons were simulated over 10 items. The degree of measurement variance τ increased across items, as it did in the first simulation study. The lower bound was $-1/m.$ Here, the lowest possible value for measurement variance would be $-1/50.$ The fractional Bayes factors were computed to detect evidence in favor of the measurement invariance hypothesis H_0 when the alternative hypotheses are no measurement invariance H_1 and measurement variance H_2.

The number of MCMC iterations was 5,000 with a burn-in of 1,000. The convergence and autocorrelation plots, created using the R package coda (Plummer et al., 2006) didn't show any irregularities. As in the previous study, this study consisted of 50 data replications, which led to stable results. The mean results from these replications are presented.

Results

Table 2 presents the results of the simulation study. In this simulation, measurement variance increases across items as in the previous simulation. The same symbols are used, where τ represents the simulated degree of measurement variance and τ' represents the measurement variance detected by the marginalized item response model. Column $\tau-\tau'$ shows the difference between the simulated degree of measurement variance and the degree of measurement variance detected by the model. The estimated degree of measurement variance τ' differs only a small amount from the simulated degree of measurement variance τ. The smallest difference is 0.000; the greatest absolute difference is 0.032. There appears to be an overestimation of the degree of measurement variance when the difference is less than 0.075 and an underestimation of the degree of measurement variance when the difference is greater than 0.125. However, this under- and overestimation is present to a lesser extent compared to the estimates for the fixed number of groups (see section titled Simulation Study for Fixed Groups). In this case, the variance in item responses between groups is also used to estimate τ. Again, the posterior mean will overestimate the true value when the posterior distribution is right-skewed and underestimate the true value when it is left-skewed.

In order to test whether $\tau = 0$ (H_0) or $\tau \neq 0$ (H_1), fractional Bayes factor FBF_{01} was computed. When looking at the natural logarithm of the fractional Bayes factor in column $\ln(FBF_{01})$ in Table 2, it can be seen that the greater the simulated degree of measurement variance τ gets, the more negative the natural logarithm of the fractional Bayes factor becomes. The FBF_{01} results show correctly that for item 1 and 2, there is more support for the null hypothesis (H_0), and for items 3 and 10 there is substantially more support for the alternative hypothesis (H_1). When looking at the results of FBF_{02} in the last column of Table 2, where the alternative hypothesis is restricted to measurement variance (H_2: τ >0), it can be seen that there is a large degree of support for the measurement invariance hypothesis for items 1 and 2. For item 3, it is approximately 5.13 times more likely that the item is measurement variant than that it is measurement invariant. The results show that items 4–10 are measurement variant. It can be concluded that the results are exactly in line with those for the simulated data.

TABLE 2

Results of the simulation study, replicated 50 times, for estimating the degree of measurement variance

Item	τ	τ'	$\tau-\tau'$	$\ln(FBF_{01})$	FBF_{01}	$\ln(FBF_{02})$	FBF_{02}
1	–0.020	0.002	–0.022	0.449	1.566	4.641	103.607
2	0.000	0.013	–0.013	–0.471	0.625	2.971	19.506
3	0.025	0.036	–0.011	–4.520	0.011	–1.636	0.195
4	0.050	0.055	–0.005	–9.415	<0.001	–6.594	0.001
5	0.075	0.075	0.000	–15.436	<0.001	–12.554	<0.001
6	0.100	0.100	0.000	–22.808	<0.001	–19.975	<0.001
7	0.125	0.125	0.000	–31.127	<0.001	–28.283	<0.001
8	0.150	0.140	0.010	–36.227	<0.001	–33.376	<0.001
9	0.175	0.143	0.032	–37.435	<0.001	–34.570	<0.001
10	0.200	0.188	0.012	–53.049	<0.001	–50.185	<0.001

FBF_{01} = fractional Bayes factor, where H_0: τ =0 and H_1: $\tau \neq 0$; FBF_{02} = fractional Bayes factor, where H_0: τ =0 and H_2: τ >0.

Compared to the simulation study with fixed groups, the fractional Bayes factor results for items 1 and 2 show more support for the measurement invariance hypothesis. There is more information in the data about the exact value of the parameter, since both within- and between-group variation is used. For fixed groups, only within-group information is used. As a result, estimates for the degree of measurement variance for randomly selected groups is more accurate compared to estimates for the degree of measurement variance for fixed groups.

Furthermore, for items 1 and 2, there is more data evidence in favor of the null hypothesis (H_0), which makes the alternative hypothesis (H_1) less attractive. Note that the data still provides some support for values near zero, which makes H_1 slightly more attractive than H_0, leading to a fractional Bayes factor $FBF_{01} < 1$ for item 2. However, when the alternative hypothesis (H_2) is considered, then small negative values of τ do not contribute to the evidence against the null hypothesis.

Evaluating Measurement Invariance Assumptions of the European Social Survey Items

There are many areas where methods for the detection of measurement variance can be useful. International surveys, in which the answers of respondents across countries are compared, are one such example. To demonstrate the application of the fractional Bayes factor under a marginalized item response model for detecting measurement variance, data from the European Social Survey (ESS) was used from round 7, year 2014, of the European Social Survey (2014). The data used for this example contains groups of different sizes. Currently, the marginalized item response model is only applicable to data with equal group sizes. In order to accomplish equal group sizes, a balanced random sample is drawn from the countries included in this empirical example.

In the case of ESS data, attitude is measured instead of ability. The interpretation of measurement variance changes slightly when attitude is measured. When measurement variance is present, this means that people from different countries who have the same attitude toward immigrants have a different probability of scoring on the negative side of the scale with respect to attitude. Otherwise stated, for a measurement variant item, respondents who have the same attitudes but are from different countries have unequal probabilities of scoring positively toward immigrants. The conclusion remains the same: When the items show measurement variance, scale scores constructed under the assumption of measurement invariance cannot be compared across countries.

In order to show the application of the model to empirical data, eight items were selected from the ESS survey (European Social Survey, 2014). The eight items selected (Table 3) concerned the topic of immigration, since it is likely that measurement variance is present in items such as these. The items contributed to the same scale, which measures attitude toward immigrants. Measurement variance was tested for two different situations: fixed number of groups and randomly selected groups. In the situation of fixed groups, the goal was to make inferences about the degree of measurement variance between two countries, in this case Belgium and Sweden. The number of observations included was 1,750 for each country. So the total number of observations was 3,500. In the situation of randomly selected groups, six countries were selected and included in the study, and the goal was to investigate measurement invariance assumptions for items across the countries included in the ESS. The six countries selected were Austria, Belgium, Switzerland, Czech Republic, Germany, and Denmark. The number of observations included was 1,500 for each country, for a total number of observations of 9,000. In the current model, additional sampling weights were not taken into account. Therefore, it is possible that the empirical results were affected by exclusion of the weights.

In order to decide whether or not measurement variance was present in an item, fractional Bayes factors were computed. As in the simulation study, the FBF_{01} represents the fractional Bayes factor to evaluate the evidence in favor of measurement invariance (H_0: $\tau = 0$) compared to no measurement invariance (H_1: $\tau \neq 0$). The FBF_{02} represents the evidence in favor of measurement invariance compared to measurement variance (H_2: $\tau > 0$).

TABLE 3
Statements of the ESS selected for the application study (European Social Survey, 2015)

Item	Statement and scale
1.	Immigrants generally take jobs away or help to create new jobs 0 Take jobs away – 10 Create new jobs
2.	Immigrants take out more than they put in regarding taxes and welfare or not 0 Generally take out more – 10 Generally put in more
3.	Immigrants make country's crime problems worse or better 0 Crime problems made worse – 10 Crime problems made better
4.	Mind if immigrant of different race or ethnic group was your boss 0 Not mind at all – 10 Mind a lot
5.	Mind if immigrant of different race or ethnic group would marry close relative 0 Not mind at all – 10 Mind a lot
6.	The country's cultural life is undermined or enriched by immigrants 0 Cultural life undermined – 10 Cultural life enriched
7.	Immigration is bad or good for country's economy 0 Bad for the economy – 10 Good for the economy
8.	Immigrants make the country a worse or better place to live 0 Worse place to live – 10 Better place to live

The data from the ESS study were dichotomized in order to make them suitable for the marginalized item response model. The possible answers to each of the included questions are on a scale from 0 to 10, as illustrated in Table 3. In items 1–3 and item 6–8, 0 stands for a negative attitude toward immigrants and 10 stands for a positive attitude toward immigrants. The five most negative categories toward immigrants (category 0–4) are coded as 1 and the other six categories (5–10) which reflect (relatively) positive attitudes about immigrants are coded as 0. For items 4 and 5, 0 stands for a positive attitude toward immigrants and 10 stands for a negative attitude toward immigrants. Again, the five most negative categories with respect to attitude toward immigrants (categories 6–10) are coded as 1 and the other six categories (categories 0–5) which reflect (relatively) positive attitudes toward immigrants are coded as 0.

The discussed MCMC algorithms were used to estimate the degree of measurement variance for the fixed and randomly selected groups, respectively. The number of iterations was 5,000 with a burn-in of 1,000. The convergence and autocorrelation plots, created using the R package coda (Plummer et al., 2006), showed no irregularities.

Table 4 shows the results for the situation where the degree of measurement variance is estimated for both fixed and randomly selected groups. First, the results for the degree of measurement variance for a fixed number of groups (i.e., Belgium and Sweden) will be discussed. Results for the fractional Bayes factors FBF_{01} and FBF_{02} are presented in this table as well as the results for the ppp χ^2_{MH}.

From the results it can be concluded that, according to the fractional Bayes factors, none of the eight items appear to be measurement invariant. The support in favor of measurement variance is lowest for item 6, where the FBF_{02} estimate shows just around 3.94 (1/.254) more support for H_2 compared to H_0. Item 6, which concerns the question of whether the country's cultural life is undermined or enriched by immigrants, shows the lowest support for measurement variance. The item with the highest degree of measurement variance appears to be item 3, where τ is estimated to be 0.149. For this item, respondents were asked their opinion with respect to the country's crime problems. For the other six items, a large degree of support was found in favor of measurement variance, with τ' ranging from .036 to 0.080.

A discrepancy can be observed between the results for the fractional Bayes factors FBF_{01} and FBF_{02} and the results for the ppp χ^2_{MH}. The latter appears to indicate that for all items there is a substantial discrepancy between the model (in which measurement invariance is assumed) and the observed data. The most noticeable difference between the result for the fractional Bayes factor and the result for the ppp χ^2_{MH} is present for item 2. The FBF02 indicates that it is approximately 30 times more likely that item 2 is measurement variant than not: the ppp χ^2_{MH} is just higher than .05, providing some evidence that the item might not be measurement invariant. With a strict cutoff value of .05, the conclusion would be that there is no evidence that the measurement invariance hypothesis (H_0) should be rejected, which is in contrast with the conclusion based on the results for the fractional Bayes factors.

TABLE 4

Results for estimating the degree of measurement variance for items from the ESS (European Social Survey, 2014)

	Fixed Groups						Random Groups				
Item	τ	ln(FBF_{01})	FBF_{01}	ln(FBF_{02})	FBF_{02}	ppp χ^2_{MH}	τ	ln(FBF_{01})	FBF_{01}	ln(FBF_{02})	FBF_{02}
1	0.080	−18.870	<0.001	−18.870	<0.001	0.000	0.075	−123.900	<0.001	−119.325	<0.001
2	0.036	−3.569	0.028	−3.509	0.030	0.053	0.047	−79.322	<0.001	−74.747	<0.001
3	0.149	−74.763	<0.001	−74.763	<0.001	0.000	0.184	−323.815	<0.001	−319.239	<0.001
4	0.055	−6.244	0.002	−6.231	0.002	0.000	0.031	−49.707	<0.001	−45.132	<0.001
5	0.069	−8.791	<0.001	−8.791	<0.001	0.000	0.040	−63.955	<0.001	−59.379	<0.001
6	0.022	−1.772	0.170	−1.369	0.254	0.022	0.049	−82.171	<0.001	−77.596	<0.001
7	0.058	−12.931	<0.001	−12.934	<0.001	0.000	0.080	−140.381	<0.001	−135.806	<0.001
8	0.073	−20.636	<0.001	−20.636	<0.001	0.002	0.042	−67.232	<0.001	−62.656	<0.001

FBF_{01} = fractional Bayes factor where H_0: τ =0 and H_1: $\tau \neq 0$; FBF_{02} = fractional Bayes factor where H_0: τ =0 and H_2: τ >0; ppp χ^2_{MH} = the posterior predictive *p*-value based on the Mantel–Haenszel χ^2_{MH} statistic.

For randomly selected groups, measurement variance was assessed by using data from the countries Austria, Belgium, Switzerland, Czech Republic, Germany, and Denmark. Although the estimated degree of measurement variance differed strongly between items, it was remarkable that each of the eight items showed a large degree of support for the measurement variance hypothesis over the measurement invariance hypothesis. The item with the highest degree of measurement variance was item 3, with an estimated τ' of .184. This item also appeared to have the highest degree of measurement variance when only Belgium and Sweden were compared, as in the situation of fixed groups. The other seven items were considered moderately measurement variant, with estimated τ' values ranging from 0.031 to 0.080.

Conclusion and Discussion

The goal of this study was to propose a new method for detecting measurement variance using a marginalized item response model. This method uses the additional correlation between observations in order to detect the presence of measurement variance without conditioning on group-specific item parameters. That is, one common item difficulty parameter that applies to all groups is modeled. As a result, any group-specific deviations are included in the errors. Subsequently, measurement variance can be detected by evaluating the correlation between residuals within a group. The functioning of this proposed method for the detection of measurement variance is evaluated with simulation studies and applied to empirical data.

The simulation studies showed that this new method is able to estimate the degree of measurement variance for both randomly selected and fixed groups. The fractional Bayes factor was able to accurately determine whether the estimated degree of measurement variance was equal to or greater than zero, and it outperformed the ppp χ^2_{MH}. The results for the randomly selected groups were more convincing compared to the results for the fixed groups, because both within-group and between-group information was used in evaluating the level of measurement variance in the randomly selected groups.

For fixed groups when measurement invariance was assumed, the data showed support for parameter values around zero for the specified simulated conditions, which led to slightly more support for the alternative hypothesis of no measurement invariance (H_1). The fractional Bayes factor was less than one but did not show significant support for H_1. When the alternative hypothesis was specified to be measurement variance (H_2), a large degree of support in favor of the measurement invariance hypothesis (H_0) was found under simulated conditions.

The posterior mean is used as a point estimator of the covariance parameter, which has a skewed posterior distribution. When measurement variance is relatively low and the distribution is right-skewed, the posterior mean tends to overestimate the degree of measurement variance. When the degree of measurement variance is relatively high and the distribution is left-skewed, the posterior mean tends to underestimate the degree of measurement variance. This is a property of the posterior mean as a point estimator and does not relate to the properties of the proposed fractional Bayes factors, whose computations are based on sampled values from the posterior, taking into account any skewness of the posterior.

This report shows that the model can be applied to empirical data. Data regarding the attitude toward immigration questioned in the ESS was used to illustrate the method. The results show that measurement variance appears to be present in all items included in this empirical example. Further developments are needed to make the method applicable to unequal group sizes.

A limitation of the new method is that it can only be applied to dichotomous data. The extension to polytomous data would make the method for the detection of measurement variance more widely applicable. This generalization would require a data augmentation scheme for polytomous items, as described in Fox (2010, Chapter 7). Subsequently, the covariance structure of augmented responses to each response category needs to be evaluated to evaluate the measurement invariance assumptions of the threshold parameters of the items. International studies often contain person weights; it would also be a practical addition to include weights in the analysis. The likelihood could be weighted given sampling weights. This would lead to a posterior from which samples cannot be directly drawn. A Metropolis-Hastings algorithm could be used to draw samples from the posterior distribution to compute the fractional Bayes factor.

References

Azevado, C., Andrade, D., & Fox, J.-P. (2012). A Bayesian generalized multiple group IRT model with model-fit assessment tools. *Computational Statistics and Data Analysis, 56*, 4399–4412.

Bock, R., & Zimowski, M. (1997). *The multiple groups IRT*. Springer-Verlag. In W. J. Van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory.* New York: Springer-Verlag.

De Jong, M., Steenkamp, J., & Fox, J.-P. (2007). Relaxing measurement invariance in cross-national consumer research using a hierarchical IRT model. *Journal of Consumer Research, 34*, 260–278.

European Social Survey. (2014). *Round 7 data. Data file edition 1.0.* Norwegian Social Science Data Services, Norway—Data Archive and distributor of ESS data for ESS ERIC.

European Social Survey. (2015). *ESS-7 2014 documentation report. Edition 1.0.* Bergen, European Social Survey Data Archive, Norwegian Social Science Data Services for ESS ERIC.

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications.* New York: Springer.

Fox, J.-P., Sinharay, S., & Mulder, J. (2016). Bayes factor covariance testing in item response models. *Psychometrika*. Manuscript submitted for publication.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6*, 733–807.

Hambleton, R., & Rogers, H. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel–Haenszel methods. *Applied Measurement in Education*, *2*(4), 313–334.

Holland, P., & Wainer, H. (1993). *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Kass, R., & Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773–795.

Kelcey, B., McGinn, D., & Hill, H. (2014). Approximate measurement invariance in cross-classified rater-mediated assessments. *Frontiers in Psychology, 5*(1469), 1–13.

Klugkist, I., & Hoijtink, H. (2007). The Bayes factor for inequality and about equality constrained models. *Computational Statistics & Data Analysis, 51*, 6367–6379.

Magis, D., Beland, S., & Raiche, G. (2015). difR: Collection of methods to detect dichotomous differential item functioning (DIF) [Computer software manual]. (R package version 4.6)

O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society. Series B (Methodological), 57*(1), 99–138.

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R news, 6*(1), 7–11. Retrieved from https://www.r-project.org/doc/Rnews/Rnews_2006-1.pdf

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/

Raftery, A. (1995). Bayesian model selection in social research. *Sociological Methodology, 25*, 111–163.

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York: Wiley.

Sinharay, S., Johnson, M., & Stern, H. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement, 30*, 298–321.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin, 99*(1), 118–128.

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthen, 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*(770), 1–15.

Verhagen, J., & Fox, J.-P. (2013). Bayesian tests of measurement invariance. *British Journal of Mathematical and Statistical Psychology, 66*(3), 383–401.

Verhagen, J., Levy, R., Millsap, R., & Fox, J.-P. (2015). Evaluating evidence for a Bayes factor applied to testing measurement invariance in IRT models. *Journal of Mathematical Psychology*. Retrieved from http://dx.doi.org/10.1016/j.jmp.2015.06.005.

Appendix A: The MCMC Algorithm

In this Appendix, the posterior distributions of the sampling steps of the MCMC algorithm for the marginalized item response model are discussed. The posterior distributions of the latent response variable Z_{ijk}, the person parameter θ_{ij}, the item difficulty parameter b_k, the covariance parameter τ_k, and the hyperparameters are given. In each MCMC iteration, the mean of the person parameter can be fixed to 0 to identify the scale. Furthermore, to identify the scale for the covariance structure in Equation (6), parameter $\sigma^2_{\varepsilon k}$ is restricted to one.

The Latent Response Variable Z_{ijk}

The latent response variable Z_{ijk} is sampled from a truncated normal distribution, with $Z_{ijk} > 0$ if $Y_{ijk} = 1$, and $Z_{ijk} \leq 0$ otherwise. The responses $\mathbf{Z}_{jk}$ of group j to item k are multivariate normally distributed. Therefore, the sampling of Z_{ijk} should also take into account the other responses of group j to item k. The mean and variance of the normal distribution in Equation (8) can be obtained in closed form, using the expression for the inverse of a compound-symmetry covariance matrix (Fox, 2010, pp.151–152). Then, deriving the expression of the mean and variance of the conditional normal distribution, it follows that

$$mu_{z_{ijk}} = \theta_{ij} - b_k + \frac{\tau_k}{1+(m-1)\cdot\tau_k}\cdot\sum_{l=1,l\neq i}^{m}(Z_{ljk} - (\theta_{lj} - b_k))$$

$$var_{z_{ijk}} = \frac{1+m\tau_k}{1+(m-1)\tau_k},$$

for the covariance structure in Equation (6). For the fixed group setting with covariance structure (7), the mean and variance are given by

$$mu_{z_{ijk}} = \theta_{ij} - b_k + \frac{\tau_k}{1-\tau_k+(m-1)\cdot\tau_k}\cdot\sum_{l=1,l\neq i}^{m}(Z_{ljk} - (\theta_{lj} - b_k))$$

$$var_{z_{ijk}} = \frac{1-(m-1)\tau_k^2}{1-\tau_k+(m-1)\tau_k}.$$

The Person Parameter θ

According to Equation (9), the person parameter θ_{ij} given $\mathbf{Z}_{ij}$, $\mathbf{b}$, $\boldsymbol{\tau}$, μ_{θ_j}, and σ_θ^2 is sampled from a normal distribution with mean $mu_{\theta_{ij}}$ and variance $var_{\theta_{ij}}$, where the mean and variance are given, respectively, by

$$mu_{\theta_{ij}} = \frac{\sum_{k=1}^{K} Z_{ijk}(1+\tau_k)^{-1} + \mu_{\theta_j}\sigma_\theta^{-2}}{\sum_{k=1}^{K}(1+\tau_k)^{-1} + \sigma_\theta^{-2}},$$

$$var_{\theta_{ij}} = \frac{1}{\sum_{k=1}^{K}(1+\tau_k)^{-1} + \sigma_\theta^{-2}}.$$

Given a uniform hyperprior, parameter μ_{θ_j} is sampled from a normal distribution with mean $\bar{\theta}_j$ and variance σ_θ^2/m. The hyperparameter σ_θ^2 is sampled from an inverse gamma distribution with shape parameter $(N/2+\alpha)$ and scale parameter $(SS/2+\beta)$, where SS represents the within-group sum of squares:

$$SS = \sum_{j=1}^{J}\left(\sum_{i=1}^{m}(\theta_{ij} - \bar{\theta}_j)^2\right).$$

The Item Difficulty Parameter b

According to Equation (10), the item difficulty parameter b_k can be sampled from a conditional normal distribution, given $\mathbf{Z}_k$, τ_k, μ_b and σ_b^2, with the mean mu_{b_k} and variance var_{b_k} given, respectively, by

$$mu_{b_k} = var_{b_k}\left(\sum_{j=1}^{J}\frac{-\sum_{i=1}^{m} Z_{ijk}}{1+m\tau_k} + \frac{\mu_b}{\sigma_b^2}\right)$$

$$var_{b_k} = \left(\frac{J}{(m\tau_k+1)/m} + \frac{1}{\sigma_b^2}\right)^{-1}.$$

The hyperparameter μ_b is sampled from a normal distribution with mean

$$\mu = \left(\frac{K\sigma_b^2}{\sigma_b^2(K+n_0)} \right) \overline{b},$$

where $\overline{b}$ is the mean item difficulty across items, and variance

$$\sigma^2 = \frac{1}{\sigma_b^2(K+n_0)}.$$

Hyperparameter σ_b^2 is sampled from an inverse gamma distribution with the shape parameter equal to $(K+\alpha_b)/2$ and the scale parameter equal to $SS/2,$ with

$$SS = \beta_b 1 + \sum_{k=1}^{K} \left(b_k - \overline{b}\right)^2 + \frac{Kn_0}{2(K+n_0)} \overline{b}.$$

The Degree of Measurement Variance τ

The within-group sum of squares and the between-group sum of squares specify both levels of variability in item responses. The within-group sum of squares is defined as

$$S_{w_k} = \sum_{j=1}^{J} \left(\sum_{m=1}^{M} \left(\overline{E}_{jk} - E_{ijk} \right)^2 \right), \quad \text{(A-1)}$$

where $E_{ijk} = Z_{ijk} - (\theta_{ij} - b_k)$ is the response error and $\overline{E}_{jk}$ the average error in group j concerning responses to item $k.$ The between-group sum of squares is defined as

$$S_{b_k} = \sum_{j=1}^{J} (\overline{E}_{jk} - \overline{E}_k)^2, \quad \text{(A-2)}$$

where $\overline{E}_k$ is the average latent response error of item $k.$

For the covariance structure defined in Equation (6), Fox et al. (2016) showed that the $\tau_k + m^{-1}$ is sampled from an inverse-gamma distribution with shape parameter $(J-1)/2$ and scale parameter $S_b/2$.

For the covariance structure defined in Equation (7), following Fox et al. (2016), let $\widetilde{\boldsymbol{Z}}_{jk} = \boldsymbol{HE}_{jk}$, where $\boldsymbol{H}$ is the orthogonal Helmert matrix. The mean and variance of the transformed first components $\widetilde{Z}_{1jk}$ is equal to 0 and $1+(m-1)\tau_k$, respectively (Rao, 1973, pp. 196–197). Given the prior in Equation (12), the posterior distribution of $(m-1)^{-1}+\tau_k$ is given by

$$p(\tau_k \mid \tilde{z}_{1k}) \propto \left((m-1)-1+\tau_k\right)^{-J/2-1} \exp\left(\frac{\frac{-m}{m-1}\sum_{j=1}^{J}\left(\tilde{z}_{1jk}\right)^2 / 2}{(m-1)^{-1}+\tau_k}\right)$$

$$\propto \left((m-1)-1+\tau_k\right)^{-J/2-1} \exp\left(\frac{\frac{-m}{m-1}\sum_{j=1}^{J}\left(\bar{z}jk-(\bar{\theta}_j-b_k)\right)^2 / 2}{(m-1)^{-1}+\tau_k}\right)$$

$$p(\tau_k \mid \tilde{z}_{jk}) = \frac{\left(\widetilde{S}_b / 2\right)^{J/2}}{\Gamma(J/2)}\left((m-1)^{-1}+\tau_k\right)^{-J/2-1} \exp\left(\frac{-\widetilde{S}_b / 2}{(m-1)^{-1}+\tau_k}\right), \qquad \text{(A-3)}$$

where $\widetilde{S}_b = \frac{m}{m-1} S_b / 2.$ The normalizing constant follows from the inverse gamma distribution. As a result, for the covariance structure in Equation (7), the $\tau_k + 1/(m-1)$ can be sampled from an inverse-gamma distribution with shape parameter $J/2$ and scale parameter $\widetilde{S}_b$.

The remaining transformed components $\widetilde{Z}_{2jk},...,\widetilde{Z}_{mjk}$ are independently and identically normally distributed with mean 0 and variance $\sigma^2_{\varepsilon k} = 1-\tau_k$. Given the prior in Equation (13), the posterior distribution of $\sigma^2_{\varepsilon k}$ is given by

$$p(\sigma^2_{\varepsilon k} \mid \tilde{z}_{jk}) \propto \left(\sigma^2_{\varepsilon k}\right)^{-J(m-1)/2-1} \exp\left(\frac{-\sum_{j=1}^{J}\sum_{i=2}^{m}\left(\tilde{z}_{ijk}\right)^2 / 2}{\sigma^2_{\varepsilon k}}\right)$$

$$\propto \left(\sigma^2_{\varepsilon k}\right)^{-J(m-1)/2-1} \exp\left(\frac{-\sum_{j=1}^{J}\sum_{i=2}^{m}\left(z_{ijk}-\bar{z}_{jk}\right)^2}{2\sigma^2_{\varepsilon k}}\right)$$

$$p(\sigma_{\varepsilon k}^2 \mid \tilde{z}_{jk}) = \frac{(S_w/2)^{J(m-1)/2}}{\Gamma(J(m-1)/2)} (\sigma_{\varepsilon k}^2)^{-J(m-1)/2-1} \exp\left(\frac{-S_w/2}{\sigma_{\varepsilon k}^2}\right), \tag{A-4}$$

where the normalizing constant has been obtained by recognizing that the posterior of $\sigma_{\varepsilon k}^2$ is an inverse gamma distribution. It follows that the variance component $\sigma_{\varepsilon k}^2$ can be sampled from an inverse gamma distribution with shape parameter $J(m-1)/2$ and scale parameter $S_w/2$. With $\sigma_{\varepsilon k}^2 = 1 - \tau_k$, the possible values sampled from the inverse gamma distribution are restricted to the set [1/2,1]. Subsequently, by computing the intraclass correlation coefficient from the sampled values, the sampled values for τ_k can be obtained.

Appendix B: The Fractional Bayes Factor to Test Measurement Invariance

In order to test whether measurement variance is actually present or not, a fractional Bayes factor is computed, where the probability of the data given that $\tau = 0$ (H_0) is tested against the probability of the data given that $\tau \neq 0$ (H_1), and that $\tau > 0$ (H_2). Instead of computing the traditional Bayes factor, we computed the fractional Bayes factor as described by O'Hagan (1995) and Fox et al. (2016) in order to accommodate the use of improper priors.

A minimal information sample is used with the purpose of normalizing the data under the hypothesis. In order to take into account the improper prior, the marginal distribution of the data given the hypothesis is divided by the marginal distribution of the data taken to a power denoted by s (Fox et al., 2016). Here, s symbolizes the minimal information needed to take into account the improper prior:

$$p\left(\mathbf{y} \mid \mathrm{H_i}, s\right) = \frac{\int p\left(\mathbf{y} \mid \omega_i, \mathrm{H_i}\right) p\left(\omega_i \mid \mathrm{H_i}\right) d\omega_i}{\int p\left(\mathbf{y} \mid \omega_i, \mathrm{H_i}\right)^s p\left(\omega_i \mid \mathrm{H_i}\right) d\omega_i}, \tag{B-1}$$

where, ω_i stands for the parameters in the model under hypothesis H_i.

For the covariance structure defined in Equation (6), the improper prior in Equation (12) is used. Assume a total of N responses to item k, and a balanced design for J groups with each m group member. Following Fox et al. (2016) and using the fact that $\tau_k + m^{-1}$ has an inverse-gamma distribution, with $s = 1/J$ the fractional Bayes factor is given by

$$\begin{aligned} BF^F{}_{01} &= \frac{p(\boldsymbol{y}_k \mid \tau_k = 0, s = 1/J)}{\int_{\frac{-1}{m}}^{\infty} p(\boldsymbol{y}_k \mid \tau_k, s = 1/J) p(\tau_k) d\tau_k} \\ &= \frac{p(z_k \mid \tau_k = 0, s = 1/J) dz_k}{\int \int_{\frac{-1}{m}}^{\infty} p(z_k \mid \tau_k, s = 1/J) p(\tau_k) d\tau_k dz_k} \\ &= \frac{\Gamma(1/2)}{\Gamma(J/2)} \int \frac{\exp\left(-2^{-1}(mS_b(1-J^{-1}))\right)}{(mS_b/2)^{-J/2}(mS_b/(2J))^{1/2}} dz_k, \end{aligned} \tag{B-2}$$

where S_b is defined in Equation (A-2). The $\mathrm{BF^F}_{01}$ given the hyperparameters and the latent response data z_k is computed in each MCMC iteration. The mean estimate across MCMC iterations is an estimate of the final fractional Bayes factor.

When the alternative hypothesis represents τ_k >0, the BF in Equation (B-2) is represented by

$$BF^F{}_{02} = \frac{p(z_k \mid \tau_k = 0, s = 1/J)dz_k}{\int \int_0^\infty p(z_k \mid \tau_k, s = 1/J)p(\tau_k)d\tau_k dz_k}$$
$$= \int \frac{\exp\left(-2^{-1}(mS_b(1-J^{-1}))\right)}{C_1(1-F(1/m, J/2, S_b/2))/(1-F(1/m, 1/2, S_b/(2J)))} dz_k,$$

where *F()* denotes the inverse-gamma cumulative distribution function of $\tau_k + m^{-1}$ and $F(1/m, J/2, S_b/2)$ the cumulative probability that $\tau_k + m^{-1}$ is less than m^{-1} (i.e., $\tau_k \le 0$), given a shape parameter of *J/2* and a scale parameter of $S_b/2$. The C_1 is a necessary correction for the normalizing constant of the inverse-gamma function which is equal to

$$C_1 = \frac{\Gamma(J/2)/(S_b/2)^{J/2}}{\Gamma(1/2)/(S_b/2J))^{1/2}}.$$

For the hypothesis $\tau_k > 0$, the computation of the marginal distribution of the data involves an integration of τ_k over a subset of the parameter space, and this complicates the expression for the BF.

For the covariance structure defined in Equation (7), improper priors are used for parameters τ_k and $\sigma^2_{\varepsilon k}$; see Equations (12) and (13), respectively. In the fractional Bayes factor, $s_1 = N_1^{-1} = 1/(J(m-1))$ and $s_2 = 1/J$ are defined as minimum information to deal with the improper priors. Under the null hypothesis, representing measurement invariance for item k, τ_k =0 and $\sigma^2_{\varepsilon k}$ =1. Under the alternative hypothesis, parameter $\sigma^2_{\varepsilon k}$ is defined on the interval $[1/2,1]$, referred to as H_{u_σ}; parameter τ_k is defined on $\left[-1/(m-1),\ 1/2\right]$, referred to as H_{u_τ}.

Then, the fractional Bayes factor is defined as the ratio of marginal distributions of the data under both hypotheses, that is,

$$BF_{01}{}^{F} = \frac{p(y_k \mid \sigma_{\varepsilon k}^2 = 1, \tau_k = 0, s_1 = 1/N_1, s_2 = 1/J)}{\int \int p(y_k \mid \sigma_{\varepsilon k}^2, \tau_k, s_1 = 1/N_1, s_2 = 1/J)\, p(\tau_k)\, p(\sigma_{\varepsilon k}^2)\, d\tau_k\, d\sigma_{\varepsilon k}^2}$$

$$= \frac{\int m_0(z_k, s_1 = 1/N_1, s_2 = 1/J)\, dz_k}{\int m_u(z_k, s_1 = 1/N_1, s_2 = 1/J)\, dz_k}.$$

A Helmert transformation is applied to the latent response, and the marginal distribution of the transformed response data can be factorized to find the expressions for τ_k and $\sigma_{\varepsilon k}^2$. The marginal distribution of the latent response data of item *k* under the measurement invariance hypothesis is given by

$$m_0\left(z_k, s_1, s_2\right) = \frac{p(\tilde{z}_{2k}, \ldots, \tilde{z}_{mk} \mid \sigma_{\varepsilon k}^2 = 1)\, p(\tilde{z}_{1k} \mid \tau_k = 0)}{p(\tilde{z}_{2k}, \ldots, \tilde{z}_{mk} \mid \sigma_{\varepsilon k}^2 = 1)^{1/N_1}\, p(\tilde{z}_{1k} \mid \tau_k = 0)^{1/J}}$$
$$= (2\pi)^{-(N/2-1)} \exp\left(-1/2\left(S_w(1 - N_1^{-1}) + \tilde{S}_b(1 - J^{-1})\right)\right),$$

where $\tilde{S}_b = \frac{m}{m-1} S_b$, and S_w and S_b are defined in Equations (A-1) and (A-2), respectively.

For the alternative hypothesis, the marginal distribution of the latent response data to item *k* is obtained by integrating out the covariance parameters in the expressions for the Helmert-transformed data. It follows that

$$m_u(z_k, s_1, s_2) = \frac{\int_{\sigma_{\varepsilon k}^2 \in H_{u_\sigma}} p(\tilde{z}_k \mid \sigma_{\varepsilon k}^2)\, p(\sigma_{\varepsilon k}^2)\, d\sigma_{\varepsilon k}^2 \int_{\tau_k \in H_{u_\tau}} p(\tilde{z}_k \mid \tau_k)\, p(\tau_k)\, d\tau_k}{\int_{\sigma_{\varepsilon k}^2 \in H_{u_\sigma}} p(\tilde{z}_k \mid \sigma_{\varepsilon k}^2)^{1/N_1}\, p(\sigma_{\varepsilon k}^2)\, d\sigma_{\varepsilon k}^2 \int_{\tau_k \in H_{u_\tau}} p(\tilde{z}_k \mid \tau_k)^{1/J}\, p(\tau_k)\, d\tau_k}$$
$$= (2\pi)^{-(N/2-1)} \frac{\Gamma\left(\frac{N_1}{2}\right)\Gamma\left(\frac{J}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^2} \left(\frac{S_w}{2}\right)^{-\frac{N_1}{2}} \left(\frac{\tilde{S}_b}{2}\right)^{-\frac{J}{2}} \left(\frac{S_w}{2N_1}\right)^{\frac{1}{2}} \left(\frac{\tilde{S}_b}{2J}\right)^{\frac{1}{2}} \times \qquad \text{(B-3)}$$
$$\left(\frac{F\left(1; \frac{N_1}{2}, \frac{S_w}{2}\right) - F\left(\frac{1}{2}; \frac{N_1}{2}, \frac{S_w}{2}\right)}{F\left(1; \frac{1}{2}, \frac{S_w}{2N_1}\right) - F\left(\frac{1}{2}; \frac{N_1}{2}, \frac{S_w}{2N_1}\right)}\right)\left(\frac{F\left(\frac{1}{2}, \frac{J}{2}, \frac{\tilde{S}_b}{2}\right)}{F\left(\frac{1}{2}; \frac{1}{2}, \frac{\tilde{S}_b}{2J}\right)}\right),$$

where *F()* denotes the inverse-gamma cumulative distribution function. The last term of gamma probabilities follows from the truncation of the covariance parameters to the intervals

H_{u_σ} and H_{u_τ}. The BF^{F}_{01} given the hyperparameters and the latent response data z_k is computed in each MCMC iteration. The mean estimate across MCMC iterations is an estimate of the final fractional Bayes factor.

For the alternative hypothesis $\tau_k > 0$, the marginal distribution of the data in Equation (B-3) will be slightly modified, since the integration of τ_k is restricted to (0,1/2). This leads to a small modification of the cumulative inverse-gamma probability concerning parameter τ_k, and the last term on the right-hand side of Equation (B-3) becomes

$$\frac{F\left(\frac{1}{2},\frac{J}{2},\frac{\tilde{S}_b}{2}\right)-F\left(\frac{1}{m-1},\frac{J}{2},\frac{\tilde{S}_b}{2}\right)}{F\left(\frac{1}{2},\frac{1}{2},\frac{\tilde{S}_b}{2J}\right)-F\left(\frac{1}{m-1},\frac{1}{2},\frac{\tilde{S}_b}{2J}\right)}.$$