Bayesian covariance structure modelling for measurement invariance testing

Jean-Paul Fox¹ · Jesse Koops² · Remco Feskens² · Lukas Beinhauer^{1,2}

Received: 4 October 2019 / Accepted: 7 February 2020 / Published online: 14 July 2020
© The Author(s) 2020

Abstract

In a Bayesian Covariance Structure Model (BCSM) the dependence structure implied by random item parameters is modelled directly through the covariance structure. The corresponding measurement invariance assumption for an item is represented by an additional correlation in the item responses in a group. The BCSM for measurement invariance testing is defined for mixed response types, where the additional correlation is tested with the Bayes factor. It is shown that measurement invariance can be tested simultaneously across items and thresholds for multiple groups. This avoids the risk of capitalization on chance that occurs in multiple-step procedures and avoids cumbersome procedures where items are examined sequentially. The proposed measurement invariance procedure is applied to PISA data, where the advantages of the method are illustrated.

Keywords BCSM · Random item parameters · Bayesian IRT

1 Introduction

It is most important to have a common measurement scale across groups (e.g., countries, schools), when comparing scores. Results of different groups can only be compared directly, when they are measured on a common scale. Items may only be calibrated on a common scale if measurement invariance holds. When the conditional probability of answering an item correctly depends on group information, the assumption of measurement invariance is violated (Thissen et al. 1986). Measurement invariance is the key-assumption when making group comparisons and

Communicated by Kazuo Shigemasu.

✉ Jean-Paul Fox
j.p.fox@utwente.nl

¹ University of Twente, Enschede, The Netherlands

² Cito, Enschede, The Netherlands

requires serious attention in any statistical analysis on group comparisons (e.g., Thissen et al. 1986; Fox 2010, Chapter 7; Van de Vijver and Tanzer 2004).

Fox et al. (2017) proposed to model additional correlations in the response data, when conditioning on common (invariant) item parameters. Errors in an item response model are assumed to be conditionally independently distributed given common item parameters, when assuming measurement invariance. When it is necessary to condition on group-specific item parameters to obtain conditionally independently distributed errors, this assumption is violated. The general idea is that an item-group effect can be detected as an additional correlation, when the assumption of measurement invariance is violated. This innovative Bayesian method for measurement invariance testing has also been discussed in Van de Vijver et al. (2019) for binary data in which a comparison is made with the Mantel–Haenszel test.

The method is based on the random-item parameter model (e.g., De Jong et al. 2007; Verhagen and Fox 2013), in which random item parameters model a violation of measurement invariance across groups. Item parameters are defined for each group (i.e., country-specific item parameters), and a prior is defined for the relationship between the common (invariant) item parameters (i.e., international item parameters) and the group-specific ones. The variance among the group-specific parameters represents a violation of measurement invariance, when the variance is greater than zero. Only when the variance is equal to zero, measurement invariance is identified. Fox et al. (2017) proposed a marginal version of the random item effects model, where the dependence structure implied by random item parameters is directly modelled. When not conditioning on the random item parameters, errors associated with group-members' item responses are assumed to be equicorrelated. This correlation structure is modelled instead of including random item parameters in the model. A positive correlation among item responses within a group represents a violation of measurement invariance. When the correlation is zero, the item is considered to be measurement invariant. The advantage is that the correlation of zero is not a boundary value (e.g., a random effect variance of zero is a boundary value). This leads to improved testing of the assumption of measurement invariance and an improved specification of a prior that is noninformative about the measurement invariance assumption.

This approach is captured by the Bayesian Covariance Structure Model (BCSM), which includes a description of the covariance structure of the errors in the item response model. Without conditioning on group-specific parameters and only conditioning on common parameters, correlations among errors are modeled with a covariance structure model. This error component includes any group-specific deviations, when the errors are not independently distributed given the common (measurement invariant) item parameter. However, the structure of the error component is explicitly modelled through a structured covariance matrix, which makes it possible to identify violations of measurement invariance. To detect measurement variance, the correlation of group-specific errors is evaluated. Thus, additional correlation among item responses caused by violations of measurement invariance is addressed in the BCSM with a dependence structure implied by random item parameters. The BCSM approach avoids complex identification assumptions associated with the random item effects model (De Jong et al. 2007; Verhagen and Fox 2013). A further

benefit of the method is that it can be applied to both randomly selected groups (e.g., cluster sampling) and non-randomly selected groups (e.g., fixed strata) to make inferences about measurement invariance in the population and for the groups in the sample, respectively. The marginal modelling approach of the random item effects model also connects to the BCSM applications of Klotzke and Fox (2019a, b), where a covariance structure is defined to model the dependence structure in process data. Jak et al. (2013) also considered covariance components to detect violations of measurement invariance. They used a chi-square test statistic to evaluate the hypothesis of measurement invariance in a multiple-group structural equation model.

In this study, the Bayesian measurement invariance method is extended to deal with mixed item response types. The generalization to mixed response types includes the development of a multivariate probit model with a structured covariance matrix for the threshold parameters to account for group-specific deviations from the common threshold parameters. In this approach, the measurement invariance assumption of each item threshold can be evaluated simultaneously across all items using the Bayes factor. The modelling approach does not imply any additional identification restrictions. It also relaxes restriction on the sample size to obtain stable parameter estimates, since the group-specific threshold parameters do not have to be estimated.

Next, the BCSM for measurement invariance testing is discussed with a description of the marginal random item effects model as described in Fox et al. (2017). Subsequently, the modelling framework is extended to deal with random threshold parameters (Fox 2010, Chapter 7). Then, before discussing in detail the BCSM test approach, an overview is given of current Bayesian methods to test for measurement invariance. A comparison is made with the BCSM method, where shortcomings of several other Bayesian procedures are stressed. Then, the performance of the BCSM method is shown through simulation studies, and an illustration of the method is given using a real data example. The method is applied to examine the measurement invariance assumptions of PISA items in the 2015 cycle. The selected items were developed to assess students' mathematic achievements and consisted of dichotomously and polytomously scored items. The data are used to illustrate the BCSM method for a two-group and multi-group situation, and also to compare results with those of other methods. Finally, a discussion is given to show possible extensions of the method to more general situations.

2 Overview of Bayesian methods for invariance testing

When defining a random item effects model to detect measurement variance, deviations from the overall mean are specified for each group-specific item parameter (Fox 2010, Chapter 7; Kelcey et al. 2014). The between-group variance with respect to the deviations is examined in order to detect measurement variance: the higher the between-group variance the higher the degree of measurement variance. Current methods are based on a conditional modelling approach, where group-specific item parameters (e.g., random item parameters) are defined to establish conditionally local independence. Verhagen and Fox (2013) showed that the conditional Bayesian

methods can be used to test multiple invariance hypotheses for groups, which are randomly sampled from a population. They showed good power and low Type-I error rates for different sample-size conditions to detect measurement variance. However, the random item effects approach is not suitable for non-sampled groups (i.e., no cluster sampling). For a few (non-sampled) groups, the between-group variance cannot be accurately estimated. Furthermore, the random item effect variance parameter has no correct interpretation when the considered groups define the population. The well-known two-group setting is the comparison of a single focal group to a single reference group, where a grouping variable (e.g., gender or geographic location) is the subgroup-classification or stratification variable. For non-randomly sampled groups (strata), Verhagen et al. (2015) proposed another Bayes factor test, which was able to directly evaluate item difficulty parameter differences across the selected groups.

The random item effects approach have several other limitations. First, the between-group variance of the group-specific item parameters is explicitly modelled, although the object is to test whether this variance is present. The prior for the variance parameter reflects a violation of measurement invariance and cannot be considered to be noninformative. 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 et al. 2017). This complicates statistical test procedures and requires approximate methods such as an encompassing prior approach (Klugkist and Hoijtink 2007). Third, the person parameters are estimated using potentially biased item difficulty and population parameter estimates. Fourth, the above-mentioned approach of Verhagen et al. (2015) is only applicable to a non-randomly selected number of groups (strata), and the random item effects approach only to randomly selected groups (clusters), but none of the approaches is applicable to both situations.

Van de Schoot et al. (2013) specified *a priori* the variance parameter of the random item effects instead of estimating this variance parameter. In the approximate measurement invariance approach, invariant item parameters are allowed to vary across groups, where the level of variation is *a priori* defined. When the prior variance is sufficiently small, approximate measurement invariance is considered. They demonstrated that this prior for the item parameters can be used to evaluate approximate measurement invariance, and to determine acceptable differences in item functioning between groups. However, this method has several drawbacks. First, the level of possible variation in item functioning is *a priori* fixed, which is usually unknown and is actually the object of interest in measurement invariance testing. Second, the prior is centered around zero and easily favours the status of measurement invariance over the status of measurement variance. For a (single-peaked) prior centered around the mean value (e.g., a normal distribution with mean zero) the measurement invariance assumption is *a priori* favoured. Third, models with different prior variances to allow for different levels of measurement variance across items do not differ in their number of model parameters. This complicates the model selection procedure. For instance, a model with high prior variances will always be favoured by the usual information criteria (e.g., the AIC and the BIC, Kim et al. 2017; Van

de Schoot et al. (2013) in comparison to a comparable model with smaller prior variances. Fourth, the prior variance representing approximate measurement invariance depends on the sample size. In Kim et al. (2017) and Van de Schoot et al. (2013), a prior variance of .001 represents approximate measurement invariance. Davidov et al. (2015) used a variance of .05 under the approximate measurement invariance assumption. For smaller sample sizes, the approximate measurement invariance model is often selected over the true model with a prior variance of .05. As the sample size decreases, the prior variance of .001 will lead to more shrinkage of the posterior mean estimate towards the prior mean, representing approximate measurement invariance. Therefore, when sample sizes are small, the prior variance representing approximate measurement invariance, can easily represent overwhelming evidence in favour of the measurement invariance hypothesis. It is not possible to identify a specific prior variance as the allowed magnitude of variation that is acceptable as approximate measurement invariance, since the influence of the prior variance is sample dependent.

Asparouhov and Muthén (2014) introduced the alignment method to purify latent variable scores, when some of the items are not measurement invariant. In the alignment method, a rotation of the estimated solution is given, where the number of measurement invariant items is minimized. The estimated latent variable scores can be interpreted based on a subset of items that appear as measurement invariant. The alignment method has several disadvantages, when using it to identify the measurement invariant items. First, the method is not designed to evaluate the measurement invariance assumption of each single item. The method is based on re-scaling the solution of the fitted latent variable model, and single items cannot be tested with this procedure. It can only provide a set of items that appear to be measurement variant. Second, the identified items that show measurement variance highly depends on the properties of the other items in the test. Items that exhibit a small to moderate level of measurement variance might not be detected, when other items exhibit a large level of measurement variance. Third, the method does not provide tools to test formally whether the identified discrepancies in the estimated item parameters are significant. It is not clear whether multiple hypothesis testing can be employed to identify (and/or the percentage of) the measurement variant and measurement invariant items. Fourth, the rotated solution is obtained using a loss function, which assumes that there are always a few variant items and many approximately invariant items. This approach is certainly not universal and might not be realistic for large-scale assessment data. Furthermore, different loss functions will lead to different results, but it is not clear which one will be the most suitable. In general, the optimization criterion contains a priori information, which influences the final result, and this is beyond the control of the practitioner. Note that this loss function is also used in exploratory factor analysis, where it is desired to obtain subsets of small or large factor loadings. However, this approach does not translates directly to measurement invariance testing, where optimization criteria are more likely to differ across studies. Fifth, the method breaks down when more items are measurement variant than measurement invariant. However, it is not possible to verify this, since usually it is not known which items exhibit measurement variance. The alignment method returns the simplest model, where most of the items are considered to be

measurement invariant. However, this might not be the optimal and/or most realistic solution. Finally, the alignment method is restricted to multiple (fixed) groups (strata) and cannot be applied to the situation where the groups are (cluster) sampled. The method is restricted to differences across the groups in the sample.

A well-known method to test for model violations is the posterior predictive check, where a discrepancy measure is defined to evaluate a model assumption (Gelman et al. 2003). The method is straightforward, intuitive, and can be applied to assess the fit of different aspects of the model. Therefore, posterior predictive checks have been proposed to evaluate (approximate) measurement invariance. Although they are easily implemented, the method has some drawbacks. First, the interpretation of the posterior predictive p -value is complicated, since it might not be uniformly distributed under the null model. When using posterior predictive p -values to test a null hypothesis, the empirical Type-I error rates are often below nominal values (e.g., Levy et al. 2009). For this reason, posterior predictive checking is often viewed as a diagnostic measure to identify possible misfits, instead of a formal test for model misfit (e.g., Gelman et al. 2003). Second, the posterior predictive check requires a discrepancy measure. This discrepancy measure is often not completely targeted for the specific model misfit. For instance, the odds ratio (Levy et al. 2009) has been proposed as a discrepancy measure to detect violations of local independence. However, the odds ratio is a measure of association for paired observations and does not represent directly the assumption of local independence. Without an accurate representation of the model misfit by the discrepancy measure, significant results might be caused by violating another model assumption, and incorrect inferences might be drawn. Third, the posterior predictive check is mainly used to evaluate the compatibility of the model with the data and not to compare two competing models (i.e., hypotheses). However, when competing theories are investigated, and alternative hypotheses are present, the posterior predictive check is not suitable to evaluate the competing hypotheses. The problem can occur that different models, apparently reasonable in comparison to the data, lead to different results.

3 BCSM and measurement invariance testing

In the BCSM, the dependence structure is implied by random item effects. To understand the dependence structure, the random item effects model is described. Subsequently, the corresponding dependence structure is derived and explicitly modelled through a structured covariance matrix. Consider the two-parameter multilevel IRT model with random item parameters for binary item responses (Verhagen and Fox 2013). The probability of answering an item correctly is given by

$$\begin{aligned} P(Y_{ijk} = 1 \mid \theta_{ij}, a_{kj}, b_{kj}) &= \Phi(a_{kj}\theta_{ij} - b_{kj}) \\ a_{kj} &\sim N(\mu_a, \sigma_{a_k}) \\ b_{kj} &\sim N(\mu_b, \sigma_{b_k}) \\ \theta_{ij} &\sim N(\mu_{\theta_j}, \sigma_{\theta}^2), \end{aligned} \quad (1)$$

with θ_{ij} the ability of person i in group j , and a_{kj} , b_{kj} the random discrimination and difficulty parameter for item k in group j , respectively. At this point, the random item effect variances σ_{a_k} and σ_{b_k} define the variation in item functioning across groups and are restricted to be positive. However, in the BCSM the variance parameters are covariance parameters and can take on negative values. Therefore, the variance parameters are not squared in the prior specifications in Eq. (1).

Continuous latent responses are assumed represented by Z_{ijk} . In the random item effects model, this latent response variable is given by

$$Z_{ijk} = a_{kj}\theta_{ij} - b_{kj} + e_{ijk}, \quad (2)$$

where $e_{ijk} \sim N(0, 1)$ and $Z_{ijk} > 0$ if $Y_{ijk} = 1$ and $Z_{ijk} \leq 0$ if $Y_{ijk} = 0$. The random item effects are integrated out in Eq. 2 to examine the dependence structure. This is accomplished by plugging the priors for the random item parameters into Equation 2. Then, it follows that

$$\begin{aligned} Z_{ijk} &= (a_k + \epsilon_{a_k})\theta_{ij} - (b_k + \epsilon_{b_k}) + e_{ijk}, \\ &= a_k\theta_{ij} - b_k + \epsilon_{a_k}\theta_{ij} + \epsilon_{b_k} + e_{ijk} \\ &= a_k\theta_{ij} - b_k + E_{ijk}, \end{aligned} \quad (3)$$

where the error components ϵ_{a_k} and ϵ_{b_k} are normally distributed with mean zero and variance σ_{a_k} and σ_{b_k} , respectively. Without conditioning on the random item effects, the responses to item k in group j are correlated. To see this, consider the covariance of two observations, i and l , to item k in group j :

$$\begin{aligned} \text{cov}(Z_{ijk}, Z_{ljk}) &= \text{cov}(E(Z_{ijk} | \epsilon_{a_k}, \epsilon_{b_k}), E(Z_{ljk} | \epsilon_{a_k}, \epsilon_{b_k})) \\ &\quad + E(\text{cov}(Z_{ijk}, Z_{ljk} | \epsilon_{a_k}, \epsilon_{b_k})) \\ &= \text{cov}(a_k\theta_{ij} - b_{kj} + \epsilon_{a_k}\theta_{ij} + \epsilon_{b_k}, a_k\theta_{lj} - b_{kj} + \epsilon_{a_k}\theta_{lj} + \epsilon_{b_k}) + 0 \\ &= \text{cov}(\epsilon_{a_k}\theta_{ij} + \epsilon_{b_k}, \epsilon_{a_k}\theta_{lj} + \epsilon_{b_k}) \\ &= \text{cov}(\epsilon_{a_k}\theta_{ij}, \epsilon_{a_k}\theta_{lj}) + \text{cov}(\epsilon_{b_k}, \epsilon_{b_k}) \\ &= \theta_{ij}\sigma_{a_k}\theta_{lj} + \sigma_{b_k}, \end{aligned} \quad (4)$$

where the random item effects are assumed to be independent. In the same way it can be shown that the variance of Z_{ijk} is equal to $\theta_{ij}^2\sigma_{a_k} + \sigma_{b_k}$. As a result, the implied dependence structure in group j , represented by the random item effects, is given by $\Sigma_{jk} = \mathbf{I}_n + \theta_j\theta_j'\sigma_{a_k} + \mathbf{J}_n\sigma_{b_k}$. The $\mathbf{J}_n$ is a matrix of dimension n with all elements equal to one and $\mathbf{I}_n$ is the identity matrix. The BCSM for dichotomous item response data can be given by

$$P(\mathbf{y}_{jk} | \theta_j, a_k, b_k, \Sigma_{jk}) = \int_{\Omega(\mathbf{z}_{jk})} \Phi_n(\mathbf{z}_{jk} | a_k\theta_j - b_k, \Sigma_{jk}) d\mathbf{z}_{jk}, \quad (5)$$

where $\Omega(\mathbf{z}_{jk}) = \{\mathbf{z}_{jk}, z_{ijk} \leq 0 \text{ if } y_{ijk} = 0, z_{ijk} > 0 \text{ if } y_{ijk} = 1\}$ defines the allowable region for the latent responses. The BCSM in Eq. (5) represents the marginalized random item effects model of Eq. (1).

In the covariance structure of the BCSM in Eq. (5), parameters σ_{a_k} and σ_{b_k} on the off-diagonal positions represent the implied covariance between latent responses due to the clustering of responses in groups. Each parameter can take on negative as well as positive values, as long as the covariance matrix Σ_{jk} is positive definite. The value $\sigma_{a_k} = 0$ and $\sigma_{b_k} = 0$ is of specific interest, since they represent the level of measurement invariance of item discrimination and item difficulty, respectively. A positive covariance implies a violation of measurement invariance, which can be modelled by a random item effect. A negative covariance means that the grouped item responses are negatively correlated. The responses are less likely to be grouped than when randomly assigned to groups.

The measurement invariance assumption is represented by a point-null hypothesis, which has a zero probability of being true. However, the mathematical (probability) model is an approximation of a real system, where the null hypothesis represents an approximately true scenario. The measurement invariance hypothesis is most likely never exactly true, but only approximately true, where the state of no violation of measurement invariance is approximated by a point-null hypothesis. It is assumed that this hypothesis represents a reasonable approximation to some small interval at zero. Furthermore, it is claimed that the hypothesis can only be assumed to be an approximation of the reality, when it is worth testing and it is assumed true when it withstands severe tests. When the data provides more evidence for the point-null hypothesis than the alternative hypothesis, it is reasonable to assume that the item is (approximately) measurement invariant.

3.1 BCSM for ordinal data

For ordinal response data, the graded response model with random item parameters is considered (Fox 2010, chapter 7). The Y_{ijk} take on values $c = 1, 2, \dots, C_k$, and the probability of a response in category c is represented by

$$\begin{aligned} P(Y_{ijk} = c \mid \theta_{ij}, \gamma_{kj,c}, \gamma_{kj,c-1}, a_{kj}) &= P(\gamma_{kj,c-1} \leq Z_{ijk} + a_{kj}\theta_{ij} \leq \gamma_{kj,c}) \\ a_{kj} &\sim N(\mu_a, \sigma_{a_k}) \\ \gamma_{kj,c} &\sim N(\gamma_{k,c}, \sigma_{\gamma_{k,c}}) \\ \gamma_{kj,c-1} &\sim N(\gamma_{k,c-1}, \sigma_{\gamma_{k,c-1}}) \\ \theta_{ij} &\sim N(\mu_{\theta_j}, \sigma_{\theta_j}^2), \end{aligned} \quad (6)$$

where the random effect thresholds obey the order restriction, $\gamma_{kj,0} < \gamma_{kj,1} < \dots < \gamma_{kj,C_k}$ and the common threshold parameters $\gamma_{k,0} < \gamma_{k,1} < \dots < \gamma_{k,C_k}$.

The variance in group-specific threshold effects is determined by the threshold-specific variance parameters. For each item k , the strength of the correlation among item

responses in the same category and cluster is determined by this variance parameter. In the BCSM, this dependence structure is directly modelled. However, a marginal version of the random item effects model for ordinal data in Equation (6) is difficult to obtain due to the order restriction on the thresholds. Therefore, consider the cumulative version of the graded response model (Fox 2010, pp 201–202), which relates to the sequential model for ordinal data (Tutz 1990, 1991) and the cumulative model for ordinal data (McCullagh 1980). The conditional cumulative probability of scoring in or below a category c is modelled. Random item parameters are included to model the variability in group-specific item parameters across groups. This leads to the cumulative graded response model with random item parameters:

$$P(Y_{ijk} \leq c \mid \theta_{ij}, a_{kj}, \gamma_{kj,c}) = P(Z_{ijk}(c) \leq \gamma_{kj,c}), \quad (7)$$

where $Z_{ijk}(c) \sim N(a_{kj}\theta_{ij}, 1)$ for $c = 1, \dots, C_k$. Consider the model for the latent response data, following from the cumulative model in Eq. (7):

$$Z_{ijk}(c) = a_{kj}\theta_{ij} - \gamma_{kj,c} + e_{ijk}, \quad (8)$$

where $Z_{ijk}(c) > 0$ if $Y_{ijk} > c$ and $Z_{ijk}(c) \leq 0$ if $Y_{ijk} \leq c$. The random item effects can be integrated out similar to the approach for dichotomous data in Eqs. (3) and (4). It follows that a multivariate cumulative graded response model can be represented as

$$\mathbf{Z}_{jk}(c) \sim N(a_k\theta_j - \gamma_{k,c}, \Sigma_{kj(c)}), \quad (9)$$

where $\Sigma_{kj(c)} = \mathbf{I}_n + \sigma_{a_k}\theta_j\theta_j' + \sigma_{\gamma_{k,c}}\mathbf{J}_n$ and the $\mathbf{Z}_{jk}(c) \in \Omega$ with $\Omega = \{\mathbf{z}_{jk}(c), z_{ijk}(c) \leq 0 \text{ if } y_{ijk} \leq c, z_{ijk}(c) > 0 \text{ if } y_{ijk} > c\}$. The model in Eq. (9) is referred to as the BCSM for ordinal data. The graded response model with measurement invariant items follows directly by restricting the covariance parameters σ_{a_k} and $\sigma_{\gamma_{k,c}}$, for all k and c , to zero.

The covariance matrix $\Sigma_{kj(c)}$ is item- and threshold specific. Therefore, a positive covariance $\sigma_{\gamma_{k,c}}$ among the responses in or below the threshold for c to item k represents a violation of measurement invariance of the common threshold $\gamma_{k,c}$. This is also known as uniform differential item functioning (uniform DIF). The item responses are not independently distributed, when conditioning on the common threshold parameter for category c and, as a result, a violation of measurement invariance is modelled. The BCSM for ordinal data models the additional correlation of each threshold parameter, which can be tested to examine each measurement invariance assumption. Furthermore, the threshold parameters can be examined simultaneously, since the covariance matrix models simultaneously all violations of measurement invariance across items and thresholds. When examining the covariance parameter σ_{a_k} , a violation of measurement invariance can be identified if this covariance parameter is greater than zero. In that case, a modification of the common discrimination parameter is supported by the data in group j , where the covariance component $\sigma_{a_k}\theta_j\theta_j'$ represents this group-specific adjustment on the regression effect of the common discrimination parameter.

4 MCMC and priors

In this study, attention is focused on violations of measurement invariance of the difficulty parameters (BCSM in Eq. (5)) and threshold parameters (BCSM in Eq. (9)). This means examining violations of uniform DIF. Non-uniform DIF is not considered in the description of the priors and MCMC algorithm (i.e., parameter σ_{a_k} is restricted to zero for all k). Fox et al. (2017) discussed an MCMC algorithm for the BCSM for binary data (Eq. (5)). Therefore, the priors and MCMC algorithm for the BCSM for ordinal data are considered (Eq. (9)). The MCMC algorithm consists of the following steps.

MCMC: Sample latent response data

Latent response data $Z_{ijk}(c)$ are sampled for $c = 1, \dots, C_k - 1$. The conditional distribution of the $Z_{ijk}(c)$ is discussed. The conditional distribution of the other latent response vectors is constructed in the same way. Let $\mathbf{Z}_{-i,jk}(c)$ denote the vector of augmented responses to item k excluding the i th response. Furthermore, for covariance matrix $\Sigma = \mathbf{I}_n + \sigma_{\gamma_{k,c}} \mathbf{J}_n$, let $\Sigma_{i,-i} = \sigma_{\gamma_{k,c}} \mathbf{1}_{n-1}'$ denote the i -th row of the covariance matrix excluding the i -th value, and let $\Sigma_{-i,-i} = \mathbf{I}_{n-1} + \sigma_{\gamma_{k,c}} \mathbf{J}_{n-1}$ denote the covariance matrix excluding row i and column i . The conditional distribution of $Z_{ijk}(c)$ given $\mathbf{Z}_{-i,kj}$ is truncated normal with mean

$$E(Z_{ijk}(c) | \cdot) = a_k \theta_{ij} - \gamma_{k,c} + \Sigma_{i,-i} \Sigma_{-i,-i}^{-1} (\mathbf{Z}_{-i,kj} - (a_k \theta_{-i} - \gamma_{k,c})),$$

and variance

$$\begin{aligned} \text{Var}(Z_{ijk}(c) | \cdot) &= \Sigma_{i,i} - \Sigma_{i,-i} \Sigma_{-i,-i}^{-1} \Sigma_{-i,i} \\ &= \frac{1 + n \sigma_{\gamma_{k,c}}}{1 + (n-1) \sigma_{\gamma_{k,c}}}. \end{aligned}$$

where the $\mathbf{Z}_{jk}(c)$ are restricted to the region defined by Ω , as defined below Eq. (9). A closed-form expression for the inverse of the covariance matrix Σ is known (Fox 2010, pp. 151–152), which simplifies the computation of the mean and variance component.

MCMC: Sample person parameters

A multilevel normal population distribution 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

$$\begin{aligned} \mu_{\theta_j} &\sim N(\mu, \sigma^2) \\ \sigma_{\theta}^2 &\sim IG(\alpha, \beta), \end{aligned}$$

Let $\mathbf{Z}_{ij} = (Z_{ij1}(1), \dots, Z_{ij1}(C_1 - 1), Z_{ij2}(1), \dots, Z_{ijK}(C_K - 1))$, which is normally distributed with mean $\tilde{\mathbf{a}} \theta_{ij} - \boldsymbol{\gamma}$ and diagonal covariance matrix Λ with diagonal elements $(1 + \sigma_{\gamma_{1,1}}, \dots, 1 + \sigma_{\gamma_{K,C_K}})$ and $\tilde{\mathbf{a}} = \mathbf{1}_K \otimes \{\mathbf{1}_{C_k} \otimes a_k\}$. The posterior distribution of θ_{ij} is normal with variance

$$\text{Var}(\theta_{ij} \mid \cdot) = \Sigma_{\theta} = \left((\tilde{\mathbf{a}}' \Lambda^{-1} \tilde{\mathbf{a}})^{-1} + \sigma_{\theta}^{-2} \right)^{-1} \quad (10)$$

and mean

$$E(\theta_{ij} \mid \cdot) = \Sigma_{\theta} \left(\tilde{\mathbf{a}}' \Lambda^{-1} (\mathbf{Z}_{ij} + \boldsymbol{\gamma}) + \mu_{\theta_j} / \sigma_{\theta}^2 \right). \quad (11)$$

MCMC: sample threshold and discrimination parameters

The $\mathbf{Z}_{jk}(c)$ are normally distributed with mean $a_k \theta_j - \gamma_{k,c}$ and covariance matrix $\Sigma_{k,c}$. The prior for the common threshold parameter is assumed to be normal with mean μ_{γ} and variance σ_{γ}^2 . The prior for the discrimination parameter is normal with mean μ_a and variance σ_a^2 . Let $\Sigma_{\xi} = \mathbf{I}_J \otimes \Sigma_{k,c}$ denote the covariance matrix of all latent responses scoring below or above category c of item k , $\mathbf{Z}_k(c)$, $\xi_{k,c} = (a_k, \gamma_{k,c})$, $\boldsymbol{\mu}_{\xi} = (\mu_a, \mu_{\gamma})$, $\sigma_{\xi} = (\sigma_a^2, \sigma_{\gamma}^2)$, and let $\mathbf{H} = (\boldsymbol{\theta}, -\mathbf{1}_N)$. The posterior distribution of $\xi_{k,c}$ is normal with variance

$$\text{Var}(\xi_{k,c} \mid \cdot) = \tilde{\Sigma} = \left(\left(\mathbf{H}_N' \Sigma_{\xi}^{-1} \mathbf{H}_N \right)^{-1} + \sigma_{\xi}^{-1} \right)^{-1} \quad (12)$$

and mean

$$E(\xi_{k,c} \mid \cdot) = \tilde{\Sigma} \left(\mathbf{H}_N' \Sigma_{\xi}^{-1} \mathbf{Z}_k(c) + \boldsymbol{\mu}_{\xi} \sigma_{\xi}^{-1} \right), \quad (13)$$

with $\gamma_{k,c-1} < \gamma_{k,c} < \gamma_{k,c+1}$ and $a_k > 0$.

MCMC: sample covariance parameters

The augmented data for each group, $\mathbf{Z}_{jk}(c)$, is multiplied with a Helmert matrix such that the transformed components are normally distributed, where the first component contains the information about the covariance parameter. The properties of a Helmert matrix transformation can be found in Lancaster (1965). This method has been introduced by Fox et al. (2017) to obtain the conditional distribution of the covariance parameter $\sigma_{\gamma_{k,c}}$. For J balanced groups of size n , the $\mathbf{Z}_{jk}(c)$ is normally distributed with covariance matrix $\Sigma_{kj}(c) = \mathbf{I}_n + \sigma_{\gamma_{k,c}} \mathbf{J}_n$. Let $\tilde{\mathbf{Z}}_{k1}(c)$ denote the first component of the re-scaled Helmert transformed latent response data, $\mathbf{Z}_{jk}(c) - (a_k \theta_j - \gamma_{k,c})$ ($j = 1, \dots, J$). The Helmert transformed vector contains the information about the covariance parameter $\sigma_{\gamma_{k,c}}$. The conditional distribution of $\tilde{\mathbf{Z}}_{k1}(c)$ is represented by

$$p(\tilde{\mathbf{Z}}_{k1}(c) \mid \cdot) \propto (1/n + \sigma_{\gamma_{k,c}})^{-J/2} \exp \left(\frac{-S_B^2/2}{1/n + \sigma_{\gamma_{k,c}}} \right), \quad (14)$$

where $S_B^2 = \sum_{j=1}^J \left(\bar{Z}_{jk}(c) - \bar{Z}_k(c) \right)^2$ is the between-group sum of squares.

The covariance matrix $\Sigma_{kj}(c)$ is positive-definite $\sigma_{\gamma_{k,c}} > -1/n$. An improper (non-informative) prior can be specified, which is a conjugate prior for the conditional distribution in Eq. (14).

$$p(\sigma_{\gamma_{k,c}}) \propto (1/n + \sigma_{\gamma_{k,c}})^{-1}. \quad (15)$$

The posterior distribution for $\sigma_{\gamma_{k,c}}$ is a shifted-inverse gamma with shape parameter $J/2$, scale parameter $S_B^2/2$ and shift parameter $1/n$. Values of $\tilde{\sigma}_{\gamma_{k,c}} = 1/n + \sigma_{\gamma_{k,c}}$ can be drawn from an inverse gamma distribution, $\mathcal{IG}(J/2, S_B^2/2)$, to obtain draws of $\sigma_{\gamma_{k,c}} = \tilde{\sigma}_{\gamma_{k,c}} - 1/n$.

The remaining steps of the algorithm consists of sampling values for μ_{θ_j} and σ_{θ}^2 . The sampling steps can be found in Fox (2010, Chapter 7). Values for the prior parameters μ_a , σ_a^2 , and μ_{γ} can be found in Fox (2010, Chapter 4). The model is identified by rescaling the scale of the ability parameter to a mean of zero and a variance of one in each MCMC iteration.

4.1 Fractional Bayes factor

The Bayesian measurement invariance test concerns testing the hypothesis H_0 : $\sigma_{\gamma_{k,c}} = 0$, representing measurement invariance against the hypothesis H_2 : $\sigma_{\gamma_{k,c}} > 0$, representing measurement variance. The measurement invariance assumption of each (common) threshold parameter can be tested. The hypothesis H_1 : $\sigma_{\gamma_{k,c}} < 0$ indicates that the responses in each cluster are negatively correlated, when conditioning on the common threshold parameter. In that case, respondents who consider the item to be easy are clustered with others who consider the item to be difficult. This opposite direction towards the common item difficulty leads to a negative correlation among clustered responses. This hypothesis requires more attention, but this is beyond the scope of the current study. The general hypothesis is the unrestricted hypothesis H_u : $\sigma_{\gamma_{k,c}} \neq 0$.

To determine which hypothesis is supported most by the data, the marginal distribution of the data under each hypothesis is computed. This marginal distribution of the data represents the support of the data for the hypothesis. An improper prior was used for the covariance parameters, and to avoid the dependency of the BF on unknown constants, the fractional Bayes factor approach of O'Hagan (1995) is followed. Furthermore, it has been shown that the sum of squares of the Helmert transformed latent responses provides the sufficient information concerning a covariance parameter σ_{b_k} (Fox et al. 2017) or $\sigma_{\gamma_{k,c}}$ (Eq. (14)). As a result, the Fractional Bayes factor approach discussed in Fox et al. (2017) shows the expression for the marginal distribution for each hypothesis concerning the measurement invariance assumption of a difficulty parameter (for binary data) and of a threshold parameter (for ordinal data). The expressions only require the within sum of squares and the between sum of squares of the Helmert transformed latent responses and the common cluster size as arguments. For binary response data, the latent response data are defined according to Eq. (5) and for polytomous data according to Eq. (9).

5 Simulation study

In the first simulation study, the simulated data were not generated directly under the model. Instead, the measurement variance was manipulated by manipulating the threshold parameters of the items. A lower threshold led to a higher probability of a response option in one group compared to another group, where the threshold was higher. The difference in thresholds for an item represented the violation of measurement invariance for five groups. Furthermore, the difference varied in equal steps from -1 to 1 across the 15 items. For each item, the difference in thresholds was specified between group 3 and 1 (positive direction) and group 3 and 5 (negative direction), where group 3 was assumed to experience no violation of measurement invariance and thus the common threshold parameter applied to this group. The difference in thresholds were half the size between group 3 and 2 and group 3 and 4. Responses to items with three response options were simulated. Latent variable values were sampled from a normal distribution, where the group means differed and the population mean was fixed to zero. The discrimination parameters were fixed to one. A total of five groups were considered each consisting of 100 persons.

The BCSM procedure for dichotomous items was evaluated by partitioning the polytomous data sets in separate data sets with dichotomous observations. In the second study, the polytomous data were directly analysed using the BCSM procedure for polytomous items. Each sampled polytomous data set was partitioned in two data sets, where in the first set response categories two and three were merged into one, and in the second set response categories one and two were merged. The measurement invariance assumption of threshold one was examined with the first binary data set, and threshold two with the second data set. In total 1000 data sets were considered for each condition, and reported outcomes were averaged across the 1000 data sets. The BCSM for binary data (Eq. (5)) was used to evaluate the measurement invariance assumption for category one and two for the 15 items. For each data set, the MCMC-algorithm was run for 5000 iterations, with a burn-in of 1000 iterations each. Convergence was assessed using trace plots, autocorrelation plots and diagnostics from the R package *coda*. No conflicts concerning convergence were identified.

The results are presented in Table 1. The fractional Bayes factor FBF_{0u} and FBF_{02} represent the data evidence against the null hypothesis (measurement invariance) in favour of the hypothesis $H_u: \sigma_{\gamma_{k,c}} \neq 0$ and $H_2: \sigma_{\gamma_{k,c}} > 0$, respectively. For an increasing difference in the thresholds across groups the estimated covariance $\hat{\sigma}_{\gamma_{k,c}}$ deviated more from zero. When the difference was close to zero, the covariance estimates were also close to zero. There was a direct correspondence between the estimated covariance parameters and the difference in thresholds. Note that the detected additional correlation was not simulated through a covariance matrix, but it was indirectly simulated by manipulating the thresholds across groups. The violation of measurement invariance was identified through an additional correlation in the grouped responses, when conditioning on the common threshold parameter (instead of conditioning on the group-specific threshold

Table 1 Simulation study dichotomous data: FBFs for violations of measurement invariance for thresholds 1 and 2

Difference $\gamma_{k,c}$	$\hat{\sigma}_{\gamma_{k,c}}$	$\log FBF_{01}$	$\log FBF_{02}$
<i>Threshold 1</i>			
– 1.000	1.242	– 130.259	– 127.038
– 0.862	0.885	– 90.591	– 87.369
– 0.724	0.580	– 57.008	– 53.786
– 0.586	0.340	– 31.101	– 27.876
– 0.448	0.169	– 13.498	– 10.183
– 0.310	0.059	– 3.724	0.038
– 0.172	0.017	– 0.911	3.414
– 0.034	0.014	– 0.799	3.612
0.103	0.052	– 3.346	0.573
0.241	0.137	– 10.625	– 7.177
0.379	0.306	– 27.479	– 24.229
0.517	0.539	– 52.048	– 48.826
0.655	0.868	– 87.393	– 84.172
0.793	1.248	– 128.609	– 125.387
0.931	1.800	– 188.220	– 184.998
<i>Threshold 2</i>			
– 0.931	1.199	– 125.319	– 122.098
– 0.793	0.861	– 87.965	– 84.744
– 0.655	0.570	– 55.813	– 52.591
– 0.517	0.346	– 31.659	– 28.438
– 0.379	0.174	– 13.811	– 10.558
– 0.241	0.058	– 3.389	0.262
– 0.103	0.015	– 0.723	3.588
0.034	0.013	– 0.677	3.698
0.172	0.045	– 2.565	1.304
0.310	0.140	– 10.732	– 7.389
0.448	0.292	– 25.973	– 22.745
0.586	0.509	– 49.174	– 45.952
0.724	0.768	– 77.440	– 74.218
0.862	1.104	– 114.204	– 110.982
1.000	1.517	– 159.944	– 156.722

parameters). Furthermore, all covariance parameters were jointly estimated without making use of anchor items. By modelling the additional correlation among clustered responses, a common scale was identified for the groups. The measurement scale was identified by fixing the population mean of the latent variable to zero.

The values in bold represent the situations where there was no strong evidence for the hypothesis of measurement variance (H_2). The evidence in favour of measurement invariance was relatively high, when the alternative hypothesis was H_2 . Under H_2 , although conditioning on the common threshold parameter, the grouped item

responses were still positively correlated. When the alternative was the unrestricted hypothesis H_u , negative and positive values for the covariance parameter were considered to be evidence against the null hypothesis of measurement invariance. This led to a lower support for the null hypothesis, when the negative part of the parameter space of $\sigma_{\gamma_{k,c}}$ received substantial data support. This reduced the support for the null hypothesis. This part was ignored when H_2 was the alternative hypothesis.

The data evidence for the hypotheses concerning threshold 2 was slightly smaller than those for threshold 1 for positive differences. It was likely that simulated observed values did not represent well the differences in threshold values. For negative threshold differences, the FBFs were almost similar for threshold one and two. It was also apparent that more negative (positive) differences in threshold were less (more) easier realized in the observed response values leading to less (more) data support for the alternative hypotheses. Differences in thresholds across groups were identified as additional correlations. At the limits of the latent scale, and for relatively small groups, threshold differences and/or the additional correlations representing the threshold differences, might not be realized in the simulated dichotomous values.

In the second simulation study, mixed response data were simulated using five two-option items and five three-option items. Five groups of 100 persons were considered and an additional correlation among the clustered responses was simulated ranging from .01 (item 1) to .45 (item 10). The additional correlation represented the violation of measurement invariance, since a positive covariance among clustered responses implied the support for a group-specific item threshold parameter. The latent variable values were simulated from a normal distribution allowing the groups to differ in their mean values. The measurement error variances of the groups were considered to be equal. The BCSM for polytomous data (Eq. (9)) was used to evaluate the measurement invariance assumptions for the ten items. For each condition 1000 data set were simulated, and the MCMC-algorithm was run again for 5000 iterations, with a burn-in of 1000 iterations. The measurement invariance assumption of all items were simultaneously analysed, without conditioning on knowledge of anchor items.

The results of the simulation study can be found in Table 2. The data evidence in favour of measurement variance (hypothesis H_2 under the label $\log FBF_{02}$) quickly increased for increasing values of the covariance parameter. The FBF decreased for a higher correlation among clustered observations, where an additional correlation of .056 already led to an FBF of $\exp(-3.828) \approx 1/45.97$, representing strong evidence for a violation of measurement invariance. Although the covariance increased, there was less data evidence in favour of a violation of measurement invariance for the thresholds of item 6 (polytomous item) in comparison to the difficulty parameter of item 5 (dichotomous item). This reduction in data evidence could be caused by the fact that the response data to item 6 was not only used to examine the measurement invariance property of the threshold for response category 1, but also for the threshold of response category 2. This higher complexity of the polytomous response data led to a reduction in data evidence in comparison to the response data for dichotomous item 5, which only represented a single covariance structure. It was concluded that the structure of the polytomous response data was more complex than that of

Table 2 Simulation study for mixed data: FBFs for violations of measurement invariance for dichotomous and polytomous data

Item (option)	$\sigma_{\gamma_{k\alpha}}$	Mean	SD	MSE	Bias	$\log FBF_{0u}$	$\log FBF_{02}$
<i>Dichotomous items</i>							
1	0.010	0.012	0.021	0.001	0.022	-0.737	3.589
2	0.023	0.052	0.056	0.004	0.029	-3.289	0.407
3	0.056	0.098	0.093	0.010	0.042	-7.325	-3.828
4	0.089	0.145	0.120	0.018	0.056	-11.696	-8.313
5	0.122	0.191	0.166	0.032	0.070	-16.408	-13.077
<i>Polytomous items</i>							
6(1)	0.154	0.149	0.147	0.022	-0.005	-12.673	-9.217
6(2)	0.187	0.183	0.159	0.025	-0.005	-15.545	-12.200
7(1)	0.220	0.203	0.185	0.034	-0.017	-18.157	-14.767
7(2)	0.253	0.239	0.195	0.038	-0.014	-21.374	-18.068
8(1)	0.286	0.257	0.229	0.053	-0.029	-23.763	-20.422
8(2)	0.319	0.284	0.237	0.057	-0.034	-26.177	-22.884
9(1)	0.351	0.314	0.281	0.080	-0.037	-29.856	-26.525
9(2)	0.384	0.341	0.288	0.085	-0.043	-32.181	-28.915
10(1)	0.417	0.391	0.353	0.125	-0.026	-38.206	-34.892
10(2)	0.450	0.418	0.354	0.126	-0.032	-40.556	-37.294

the dichotomous response data, which led to a reduction in data evidence in favour of the true hypothesis of measurement variance. Note that the FBFs were of comparable size when comparing the results of Table 2 with those of Table 1 for comparable covariance values. This makes sense given that the sample sizes were similar.

The estimation of the covariance parameter was considered in more detail for the mixed response data. In Table 2, the posterior mean, posterior standard deviation, mean square error and bias are given. The reported posterior mean estimate represents a trimmed mean (ignoring the 10% largest and smallest values) and was averaged across results of 1000 replications. For small covariances the posterior mean overestimated the true value, since the posterior distribution of the covariance parameter was right-skewed. As a result, the posterior mean estimates showed positive bias for smaller covariance values. For the polytomous items the detected covariance estimate was smaller than those based on dichotomous response data, most likely due to the increased complexity of the polytomous data. Furthermore, for medium-size covariance parameters the posterior distribution of the covariance parameter was less skewed leading to improved estimation results using the posterior mean estimator. The bias was much smaller and negative, representing a slight underestimation of the true covariance value.

However, for higher covariance values, the posterior mean underestimated the true value, leading to more negative bias. For higher covariance values, the level of dependencies in the clustered responses appeared to be lower than implied by the level of covariance. The estimated posterior standard deviations were relatively high (compared to the posterior mean), which showed the variation in covariance levels

across data sets. It appeared that more data sets were generated with lower correlations than with higher correlations in comparison to the true covariance value. This could be caused by the limitations of the considered measurement scale, the medium group size, and the small number of groups. Fewer groups provide less information about the level of correlation, making it more difficult to estimate accurately and precisely the level of correlation within each group. The conclusion is that the posterior mean estimator for the intra-group correlation is not very robust, where small sample sizes and skewed posterior distributions easily lead to biased estimates. However, to evaluate the assumption of measurement invariance, the information from the posterior distribution of the covariance parameter needs to be used, which is represented by the FBF, and conclusions should not be based on point estimates of the covariance parameter.

6 Real data study

The Programme for International Student Assessment or PISA is a triennial survey intended to measure the educational performance of 15-year-old students (OECD 2017a). In 2015, about 540,000 students from 73 countries and economies participated in PISA. PISA aims to measure knowledge and skills needed to participate in modern society, focusing on the core school subjects of science, mathematics and reading. Students are administered different subsets of all questions, called test versions or booklets. Student responses collected in this multi-booklet design are modelled with an IRT model. Results can be used to monitor trends in ability in various countries. Comparisons within and across countries can be made with one assessment round. However, meaningful comparisons across groups can be made only if measurement invariance holds.

The BCSM method was applied to data from the PISA 2015 survey for a two-group and a multi-group situation. Violations of measurement invariance were examined for (1) gender groups and (2) gender groups in countries, simultaneously. For the gender-group evaluation, the results of the BCSM method were compared to those of the Mantel–Haenszel test (Holland and Thayer 1988; Mantel and Haenszel 1959) and an IRT item-pair approach for evaluating measurement invariance (Bechger and Maris 2015). The Mantel–Haenszel approach was based on a log-linear model for a three-way frequency table. The performance on the item by group (boys/girls) conditional on overall performance measured by the sum score was considered. Estimates of a common conditional odds ratio were obtained, approximately χ^2 distributed, which represented the association between item score and group given the test score.

In the IRT-based item pair approach, a comparison was made between threshold differences of a pair against the same pair in the other group (and then for all pairs). Item pairs were compared instead of regarding measurement invariance to be a single-item property. The procedure started with a separate calibration of the item parameters within each group. An overall test for measurement invariance was examined, where the null hypothesis represented no violation of measurement invariance. The null hypothesis was rejected and the analyses was continued with the

evaluation of item pairs. It was investigated for each item pair if in one group a different relative difficulty was found compared to the relative difficulty of the other group, and that item pair was considered to be subject to measurement variance. Differences between item pair difficulties were tested using a Wald test and the results were summarized in a heatmap. The method is implemented in the R-package *dexter* (Maris et al. 2019).

Responses to the mathematics items administered in the first regular computer based assessment of test version 43 were used. This test version contained 22 items measuring mathematics of which three had a maximum score of two and the remaining items a maximum of one. This test version had been administered by 12,609 students within 58 countries. The BCSM method required complete observations and balanced groups. In order to end up with balanced groups, a total of 50 observations were randomly sampled within each group. The data set used in the analysis contained 5000 students from 50 different countries. In Fig. 1, the proportion correct for boys and girls are presented, as they can already serve as an indication for a different response behaviour among boys and girls.

Most of the 22 items included in the test were found below the identity line, indicating that the boys performed better than the girls. Boys had performed considerably better on item M496Q01 compared to the girls than on the other items. Contrary, item M411Q02 was one of the few items for which the proportion

Fig. 1 PISA 2015: estimated proportion correct for girls and boys on each of the 22 math items of test version 43

correct of girls was higher. This already indicates that for these items the measurement invariance assumption probably does not hold.

In Fig. 2 results of the BSCM are presented. The MCMC-algorithm was run 5000 iterations with a burn-in of 1000 iterations. Obviously, this comparison contains two groups, representing a reference group with 2500 girls and a focal group with 2500 boys. The x-axis represents the data evidence in favour of the null hypothesis of measurement invariance against the alternative hypothesis of measurement variance (hypothesis H_2 under the label $\log FBF_{02}$). The vertical dashed line was drawn at $\exp(-3.4) \approx 1/30$, representing strong evidence for a violation of measurement invariance where the measurement variance hypothesis is 30 times more likely than the measurement invariance hypothesis. It can be seen that a violation of measurement invariance was found for item M496Q01. With $\log FBF_{02} < -10$, the measurement variance hypothesis was 22,000 times more likely than the invariance hypothesis, providing decisive evidence in favour of measurement variance. For the other items no strong evidence was found for a violation of measurement invariance. Items with $\log FBF_{02} > 5$ were almost 150 times more likely to be measurement invariant than measurement variant.

Fig. 2 PISA 2015: $\log FBF_{02}$ results for all items and thresholds (first threshold $_1$ and second threshold $_2$) of measurement invariance ($\sigma_{\gamma_{k,c}} = 0$) against measurement variance ($\sigma_{\gamma_{k,c}} > 0$) for the boys and girls group

The Mantel–Haenszel (M–H) test statistics and the corresponding p -values are presented in Table 3. Items are sorted according to the p -value of the M–H test for DIF. The p -value represents the probability of observing a more extreme M–H test value than the observed one under the null hypothesis of measurement invariance. The null hypothesis was rejected, when the p -value was smaller than a significance level of .05. On the basis of the M–H test results, the assumption of measurement invariance was rejected for the highest 8 items (in bold). In comparison to the BCSM analysis, much more items were flagged as items potentially subject to DIF. This result represents a key problem in hypothesis testing using significance probabilities (e.g., Schervish 1996). The p -value depends on the sample size. For a sufficiently large sample size the M–H test results will always demonstrate a significant result given a fixed level of significance (e.g., $\alpha = .05$). Statistical significant results were obtained due to the high sample size and/or due to not adjusting the level of significance. When adjusting the level of significance, it would be possible to obtain only a significant result for item M496Q01, since that item has the highest M–H statistic value. However, in general it is not clear what the level of significance should be given a sample size. It follows from the BCSM results that most of the significant M–H results are not meaningful, and only appear to be significant due to the sufficiently large sample size. Note that given a fixed sample size, the p -values are monotonically related to the M–H statistic, and more evidence was found to reject

Table 3 PISA 2015: Mantel–Haenszel test results for gender to test the measurement invariance hypothesis

Item id	Mantel–Haenszel χ^2	p -value
M496Q01	38.708	0.000
M411Q01	22.394	0.000
M564Q02	20.000	0.000
M411Q02	19.130	0.000
M474Q01	14.881	0.000
M442Q02	14.446	0.000
M155Q02	15.293	0.000
M496Q02	4.328	0.037
M155Q04	2.987	0.084
M033Q01	2.757	0.097
M564Q01	2.495	0.114
M803Q01	2.458	0.117
M423Q01	1.637	0.201
M034Q01	1.591	0.207
M603Q01	1.447	0.229
M406Q01	0.910	0.340
M155Q03	1.949	0.377
M406Q02	0.658	0.417
M305Q01	0.410	0.522
M571Q01	0.232	0.630
M462Q01	0.511	0.775
M155Q01	0.000	0.986

the null hypothesis of measurement invariance of item M496Q01 than for instance item M411Q01. However, the order in the p -values that are higher than the significance level does not provide any information about the strength of evidence against the null hypothesis. The FBF under the BCSM represents the evidence for each hypothesis and item. The M–H test provides only evidence against the null hypothesis for those items with a p -value below the significance level.

The overall χ^2 -test for DIF between the two groups in the item-pair approach indicated that the items included in the test might be subject to DIF, ($\chi^2(63.279, df = 24)$, $p < 0.01$). The p -values associated with the item-pair differences can be used to determine whether DIF applies to the item-pair. These p -values are sample size dependent. However, the standardized item-pair differences can be considered to identify items which are potentially subject to DIF in combination with the corresponding significance probabilities. In Fig. 3, a visual summary of the standardized item-pair test results is presented. The number of items identified as potentially subject to DIF, seems much more limited compared to the M–H test. However, it remains problematic to identify the item pairs that are subject to DIF and those

Fig. 3 PISA 2015: Heatmap of standardized item-pair results

items that are not, since the significance of the test results depends on a pre-specified significance level. In the BCSM, the FBF results quantify the evidence in favour of no difference (H_0) against that of a difference (H_2) for all item differences (see for instance results of Table 1. When discussing the relevance of item-pair differences in a heatmap, it is noted that the standardized differences of item-pairs with p -values below the significance level should be considered. The standardized differences that are detected as significant need to be inspected whether these differences are indeed meaningful. When considering the most extreme results from the item-pair test, the results are in line with those from the BCSM. Again, it is item M496Q01 that can be considered as problematic and to a lesser extend this also applies to item M411Q02 (in Fig. 1 with the BCSM identified as almost strong evidence for measurement variance).

In the multi-group analysis, measurement invariance hypotheses were evaluated across countries (50 countries) and within each country for gender (boys and girls). In this multiple-group study, the BCSM approach was able to examine simultaneously all measurement invariance hypotheses for the groups of country and gender (100 groups in total). Again, the MCMC algorithm was run for 5000 iterations with a burn-in of 1000 iterations, where the FBFs were computed for each item and threshold. In Fig. 4, the BCSM FBF results are presented. The data evidence in favour of the null hypothesis of measurement invariance against the alternative hypothesis of measurement variance (hypothesis H_2 under the label $\log FBF_{02}$) is presented on the x-axis. It follows that there is substantially more evidence for measurement variance, since the number of groups is much larger. An increase in the total number of groups leads to substantially more data information. However, even for 100 groups due to conditioning on the country level, a subset of items still appeared to be measurement invariant. This showed that the increase in data information did not lead to a suspicious increase in violations of measurement invariance. An increase in evidence was found in favour of measurement variance for some items, for instance items M411Q01 and M411Q02 in line with the M–H test and the item-pair test.

The disaggregated analysis at the level of countries did lead to different results for other items. Items that appeared to be measurement invariant across the gender groups (see Fig. 1), turned out to be measurement variant when the gender groups per country were considered and vice versa. For instance, in the two-group analysis, there was strong evidence found that the item M462Q01 was measurement invariant ($\log FBF_{02} > 5$). However, in the country-gender analysis, there was decisive evidence that the item M462Q01 was measurement variant ($\log FBF_{02} < -200$). The conclusion was that the item did not show DIF for gender, when ignoring the clustering in countries. When accounting for the clustering of observations in countries, the item did show DIF for gender. Furthermore, in the two-group (aggregated) analysis the amount of evidence was overwhelming in favour of measurement variance for item M496Q01 but in the multi-group (disaggregated) analysis the evidence was in the opposite direction (i.e., in favour of measurement invariance). This might indicate a phenomenon of Simpson's paradox (e.g., Pearl 2014). In the disaggregated population, the item appeared to be measurement variant (Fig. 4), where the opposite conclusion was reached in the aggregated conditioning (Fig. 1). However, to investigate this properly, it would be necessary to test for gender DIF within each country. Then, the covariance

Fig. 4 PISA 2015: $\log FBF_{02}$ results for all items and thresholds (first threshold $_1$ and second threshold $_2$) of measurement invariance ($\sigma_{\gamma_{k,c}} = 0$) against measurement variance ($\sigma_{\gamma_{k,c}} > 0$) for 100 country-gender groups

structure needs to be country-specific, and gender DIF is allowed to vary across countries. This multiple-group BCSM approach has been discussed by Klotzke and Fox (2019b), but this is beyond the scope of the current study.

In this study, with the BCSM method additional correlations in the response data were detected, when conditioning on the common (international) item parameters. To account for the (gender and country) differences, it is possible to model the additional correlations through a structured covariance matrix, which is an efficient way to avoid the estimation of country-specific item parameters. In PISA 2015, between-country measurement variance has been taken into account by conditioning on unique national item parameter estimates (OECD 2017b). In both BCSM analyses, next to the measurement variant items, measurement invariant items were detected to define a common measurement scale. The partial measurement invariance of the instrument gives support to a common scale analysis of the PISA data.

7 Discussion

The BCSM for measurement invariance testing given mixed response data has been discussed. The method is based on testing additional correlation among clustered item responses in order to detect the presence of measurement variance without conditioning on group-specific item parameters or anchor items. Violation of measurement invariance can be detected by evaluating the correlation between residuals within each group. For mixed response data, the performance of this method for the detection of violation of measurement invariance was evaluated with simulation studies and applied to empirical data. The fractional Bayes factor results show that measurement invariance assumptions can be tested simultaneously across all items, response categories, and groups without conditioning on anchors or group-specific item parameters. This makes the procedure a very flexible and a straightforward extension of the (Probit) IRT models. The approach does not depend on a cluster sampling of groups and also applies to selected groups in the sample. The results of the real data example showed that a multi-group test method is needed. For instance, the measurement invariance assumption can be violated (across gender), when conditioning on a third grouping structure (country). The proposed BCSM approach for measurement invariance testing can be extended to account for multiple grouping structures.

As reported by Fox et al. (2017), the posterior mean is not an accurate estimator of the level of correlation, since the covariance parameter 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. Note that this is a property of the posterior mean estimator and does not relate to the properties of the proposed fractional Bayes factors, whose computations are based on the entire posterior distribution. The BCSM for Bayesian measurement invariance testing has great potential and improves current methods in different ways. The advantages are as follows:

1. All items can be tested simultaneously, a sequential procedure is not needed.
2. Dependent or nested hypotheses can be tested simultaneously. That is; full, partial or single measurement invariance hypotheses can be tested simultaneously to avoid the risk of capitalization on chance. The simultaneous evaluation of multiple measurement invariance hypotheses works in a similar way as testing a single measurement invariance hypothesis.
3. The evidence (prior and data evidence) can be quantified in favour of partial or full measurement invariance, which are usually considered null hypotheses. This is in contrast to frequentist hypothesis testing, where the null hypothesis is rejected, when significant evidence is found in favour of an alternative hypothesis.
4. Objective (uninformative) and informative priors can be specified in testing measurement invariance assumptions, where the amount of prior information can be

- fully controlled. This makes it possible to use additional information, next to the sample data, to make inferences about measurement invariance.
5. In the BF test, none of the measurement invariance hypotheses needs to be true to make valid inferences. The BF test only provides information about which hypothesis is more likely, and in the decision both hypotheses are taken into account. This is in contrast to frequentist hypothesis testing, where inferences are made by assuming that the null hypothesis is true.
 6. The BCSM test approach makes it possible to interpret results on a common scale and they are statistically comparable without needing anchor items. Factor scores can be compared across groups even when all items are identified as measurement variant.
 7. The BCSM test approach can be used with stratified groups, to make inferences about differences between groups in the sample. It can also be used for groups sampled from a population, to make inferences about the measurement invariance assumptions in the population. The complexity of the method does not increase when increasing the number of groups.
 8. The BCSM test procedure leads to exact results and does not rely on asymptotic theory. The validity of the test results does not depend on the sample size, which makes the method also usable for small data sets.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Asparouhov T, Muthén B (2014) Multiple-group factor analysis alignment. *Struct Equ Model* 21:1–14
- Bechger TM, Maris G (2015) A statistical test for differential item pair functioning. *Psychometrika* 80(2):317–340
- Davidov E, Cieciuch J, Meuleman B, Schmidt P, Algesheimer R, Hausherr M (2015) The comparability of measurements of attitudes toward immigration in the European Social Survey: exact versus approximate measurement equivalence. *Public Opin Q* 79:244–266
- De Jong M, Steenkamp J, Fox J-P (2007) Relaxing measurement invariance in cross-national consumer research using a hierarchical IRT model. *J Consumer Res* 34:260–278
- Fox J-P (2010) *Bayesian item response modeling: theory and applications*. Springer Science & Business Media, Berlin
- Fox J-P, Mulder J, Sinharay S (2017) Bayes factor covariance testing in item response models. *Psychometrika* 82(4):979–1006
- Gelman A, Carlin JB, Stern HS, Rubin DB (2003) *Bayesian Data Analysis*, 2nd edn. Chapman & Hall, Boca Raton
- Holland PW, Thayer DT (1988) Differential item functioning and the Mantel–Haenszel procedure. In: Wainer H, Braun HI (eds) *Test validity*. Lawrence Erlbaum, Hillsdale, pp 129–145

- Jak S, Oort FJ, Dolan CV (2013) A test for cluster bias: detecting violations of measurement invariance across clusters in multilevel data. *Struct Equ Model Multidiscip J* 20(2):265–282
- Kelcey B, McGinn D, Hill H (2014) Approximate measurement invariance in cross-classified rater-mediated assessments. *Front Psychol* 5(1469):1–13
- Kim ES, Cao Chunhua, Wang Y, Nguyen DT (2017) Measurement invariance testing with many groups: a comparison of five approaches. *Struct Equ Model* 24:524–544
- Klotzke K, Fox J-P (2019a) Bayesian covariance structure modelling of responses and process data. *Front Psychol* 10:1675. <https://doi.org/10.3389/fpsyg.2019.01675>
- Klotzke K, Fox J-P (2019b) Modeling dependence structures for response times in a Bayesian framework. *Psychometrika*. <https://doi.org/10.1007/s11336-019-09671-8>
- Klugkist I, Hoijtink H (2007) The Bayes factor for inequality and about equality constrained models. *Comput Stat Data Anal* 51:6367–6379
- Lancaster H (1965) The Helmert Matrices. *Am Math Mon* 72(1):4–12. <https://doi.org/10.2307/2312989>
- Levy R, Mislevy RJ, Sinharay S (2009) Posterior predictive model checking for multidimensionality in item response theory. *Appl Psychol Meas* 33:519–537
- Mantel N, Haenszel W (1959) Statistical aspects of the analysis of data from retrospective studies of disease. *JNCI J Natl Cancer Inst* 22(4):719–748
- Maris G, Bechger T, Koops J, Partchev I (2019) Dexter: data management and analysis of tests. <https://CRAN.R-project.org/package=dexter>
- McCullagh P (1980) Regression models for ordinal data. *J R Stat Soc Ser B (Methodological)* 42(2):109–127
- OECD (2017a) PISA 2015 assessment and analytical framework. OECD Publishing, Paris. <https://doi.org/10.1787/9789264281820-en>
- OECD (2017b) PISA 2015 Technical Report. OECD Publishing, Paris. <https://www.oecd.org/pisa/data/2015-technical-report/>
- O'Hagan A (1995) Fractional bayes factors for model comparison. *J R Stat Soc Ser B* 57:99–138
- Schervish M (1996) P values: what they are and what they are not. *Am Stat* 50(3):203–206. <https://doi.org/10.2307/2684655>
- Pearl J (2014) Comment: understanding Simpson's paradox. *Am Stat* 68(1):8–13
- Thissen D, Steinberg L, Gerrard M (1986) Beyond group-mean differences: the concept of item bias. *Psychol Bull* 99(1):118–128
- Tutz G (1990) Sequential item response models with an ordered response. *Br J Math Stat Psychol* 43(1):39–55
- Tutz G (1991) Sequential models in categorical regression. *Comput Stat Data Anal* 11(3):275–295
- Verhagen J, Fox J-P (2013) Bayesian tests of measurement invariance. *Br J Math Stat Psychol* 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. *J Math Psychol* 72:171–182
- Van de Schoot R, Kluytmans A, Tummers L, Lugtig P, Hox J, Muthén B (2013) Facing off with Scylla and Charybdis: a comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Front Psychol* 4(770):1–15
- Van de Vijver F, et al (2019) Invariance analyses in large-scale studies. OECD Education Working Papers, No. 201, OECD Publishing, Paris. <https://doi.org/10.1787/254738dd-en>
- Van de Vijver F, Tanzer NK (2004) Bias and equivalence in cross-cultural assessment: an overview. *Revue européenne de psychologie appliquée* 54:119–135

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.