Journal of Mathematical Psychology 72 (2016) 171–182

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Evaluating evidence for invariant items: A Bayes factor applied to testing measurement invariance in IRT models

Josine Verhagen[a,*], Roy Levy[b], Roger E. Millsap[b], Jean-Paul Fox[a]

[a] *University of Twente. Drienerlolaan 5, 7522 NB, Enschede, The Netherlands*

[b] *Arizona State University, 1151 South Forest Ave, Tempe, AZ 85281, United States*

HIGHLIGHTS

- A Bayes factor test is presented to test measurement invariance of test items.
- Bayes factors allow evaluation of evidence *in favor* of the null hypothesis of invariance.
- For the proposed test it is not necessary to indicate anchor items in advance.
- Bayes factors give more information to inform invariance decisions than traditional invariance tests.

ARTICLE INFO

Article history:
Available online 12 August 2015

Keywords:
Bayesian hypothesis testing
IRT
Latent variable models
Measurement invariance
Bayes factor
Bayesian modeling

ABSTRACT

When comparing test or questionnaire scores between groups, an important assumption is that the questionnaire or test items are measurement invariant: that they measure the underlying construct in the same way in each group. The main goal of tests for measurement invariance is to establish whether support exists for the null hypothesis of invariance. Bayesian hypothesis testing enables researchers to investigate this null hypothesis, where evidence in favor of invariance is quantified using the Bayes factor. A Bayes factor for the investigation of measurement invariance assumptions of test items for randomly selected groups was developed by Verhagen and Fox (2013a). For specific groups or measurement occasions, a different Bayes factor test is proposed here, which directly evaluates item parameter differences between groups. This test is compared to recently developed frequentist measurement invariance tests based on the Wald test in a simulation study. The close-comparison with the Wald-test performance validates the proposed Bayes factor and shows the advantages of the additional information given by the Bayes factor. Both tests are applied to the investigation of measurement invariance of a geometry test (CBASE) to illustrate the use of the Bayes factor test for measurement invariance.

1. Introduction

Increasingly, test and questionnaire administration (cognitive tests, psychological questionnaires, consumer surveys, attitude questionnaires) facilitates the comparison of scores between groups (e.g. english/spanish native, countries, ethnic groups, male/female students). To make these comparisons valid, the scores of the compared groups should be on the same scale. This can be achieved by ensuring that the measurement instrument (e.g. test, questionnaire) used to determine the scores is at least partially measurement invariant. That is; some of the test items (questions) should be measurement invariant such that persons from each group with the same true value on the measured construct have the same probability of endorsing, or correctly answering, an item or question. If, for example, the aim of a test is to compare American and Chinese students on their mathematical ability, American and Chinese students with equal math ability should have the same probability of answering each of the items correctly. If some items are more difficult for either Chinese or American students with the same mathematical ability (because of the cultural context of the question, or because the math curriculum in these countries covers different topics), these items are considered measurement variant. If measurement variance is unaccounted for, the results of group-wise comparisons of scores between groups can be biased. Also, measurement variant items are a signal to the test makers that their test does not function equally in different groups, which raises questions about the validity of the test (e.g. Borsboom, Mellenbergh, & van Heerden, 2004). Therefore, before scores are compared between groups,

[*] Corresponding author.
E-mail addresses: josineverhagen@gmail.com (J. Verhagen), Roy.Levy@asu.edu (R. Levy), g.j.a.fox@utwente.nl (J.-P. Fox).

http://dx.doi.org/10.1016/j.jmp.2015.06.005

it is crucial to test whether the characteristics of the measurement instrument are invariant.

Many methods have been developed for testing measurement invariance, also known as the absence of differential item functioning (see for an overview: Millsap, 2011). However, there are limitations of these current (frequentist) invariance tests. First, it is not possible to gather evidence in favor of the invariance hypothesis. Second, (except for the alignment method, Muthén & Asparouhov, 2013b) some invariant (anchor) items need to be identified beforehand, which means not all items can be tested for invariance simultaneously. Tests for measurement invariance are specifically aimed at collecting evidence in favor of the hypothesis of invariance. The Bayes factor is especially well suited for this purpose. Contrary to frequentist tests, which can only gather evidence to reject invariance, the Bayes factor weighs evidence in favor of both the invariance and non-invariance hypotheses. In addition, by using a different restriction on the item parameters, invariance can be evaluated for all items simultaneously, without the need for anchor items.

Verhagen and Fox (2013a) proposed a Bayes factor test based on the variance of item parameters over groups to compare the nested hypotheses of invariance and non-invariance for a relatively large number of randomly selected groups (i.e. schools, countries). This test is not suitable for a small number of not randomly selected groups of specific interest (fixed groups, i.e. males/females), however, as it requires a valid estimate of the variance of an item parameter over groups. Therefore, we will propose a different Bayes factor test to evaluate measurement invariance given a small number of fixed groups, which is easy implemented.

First, the measurement model will be explained for the situation where respondents are randomly selected from a few specific groups, and interest is in the comparison of item and latent-mean parameters for the selected groups. Then, the proposed method for Bayesian hypothesis testing and its advantages for testing measurement invariance are explained. Subsequently, simulation studies will compare the performance of the proposed Bayes factor tests with Wald tests (Woods, Cai, & Wang, 2012), as implemented in IRTPRO (Cai, Thissen, & du Toit, 2011). To show the wide use of Bayesian invariance tests and its advantages over the Wald tests, the Bayes factors for the fixed group setting will be evaluated with an example concerning geometry items from the College Basic Academic Subjects Examination (CBASE).

2. Bayesian multiple group IRT models

2.1. Introducing the measurement model

Item response theory (IRT) models are a common choice as measurement models for tests and questionnaires, especially in case of discrete responses. In a mathematics test, for example, the probability of a correct response is modeled as a function of mathematical ability and the difficulty of a test item. The mathematical ability of a person is inferred by comparing the responses on a set of test items to the difficulty of those items.

The one parameter logistic model (1PL), will be used to introduce the concepts of the Bayesian framework for measurement variance modeling. In the 1PL model, the probability of a dichotomous response of person $i = 1, \ldots, N$ on item $k = 1, \ldots, K$ is modeled as a function of the threshold or difficulty of an item k, b_k (item parameter), and the score of a person on the underlying construct being measured θ_i (person parameter):

$$P(Y_{ik} = 1|\theta_i, b_k) = \frac{e^{(\theta_i - b_k)}}{1 + e^{(\theta_i - b_k)}}. \tag{1}$$

The result is a logistic function, where the probability of a correct response is a function of the difference between the ability of a person and the difficulty of an item (Fig. 1). The model assumes unidimensionality and local independence of the item responses.

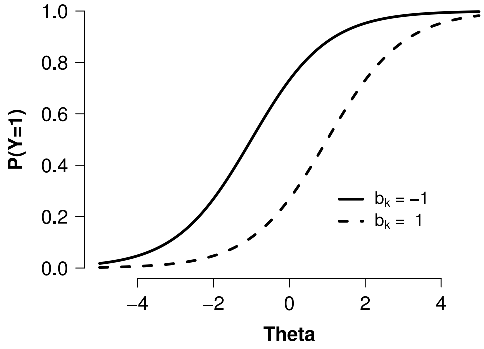

Fig. 1. An illustration of the 1PL IRT model.

2.2. Bayesian IRT models

Recently, Bayesian versions of the well-known IRT models have been developed (Albert, 1992; Fox & Glas, 2001; Patz & Junker, 1999a,b). The priors for the item parameters specify the variation among item characteristics. Bayesian IRT models known as random item effects models (e.g. De Boeck, 2008; Glas & Van der Linden, 2003; Janssen, Tuerlinckx, Meulders, & De Boeck, 2000) model the items in a test as a random sample from an item population. The prior for the item parameters is therefore specified as a normal distribution, with a common mean and variance for all items:

$$b_k \sim N(b_0, \sigma^2_{b_k}). \tag{2}$$

The posterior distributions of the separate item parameters are a combination of a function of the average percentage correct on this item averaged over all groups and the parameters for the prior distribution b_0 and $\sigma^2_{b_k}$. As a prior distribution for the person parameters a standard normal prior $N(0, 1)$ is usually chosen. More details on the estimation of Bayesian IRT models can be found in Fox (2010).

2.3. Bayesian IRT models for multiple groups

To investigate measurement invariance, it is necessary to model the responses on a test or questionnaire using an IRT model which allows group differences in both test scores and item characteristics. So-called multiple-group IRT models (for an overview, see Bock & Zimowski, 1997), in which each group has a specific latent trait distribution, can be used to account for the nesting of respondents in groups and the item characteristics that vary across groups.

In this paper, multiple group IRT models are considered with group-specific item parameters to model variation in item functioning over groups, besides the variation across items. Bayesian IRT models are easily extended in this way to form a multiple-group IRT model. Although group-specific random item parameters were originally used to model the nesting of items within testlets or item families (Bradlow, Wainer, & Wang, 1999; Glas & Van der Linden, 2003; Glas, van der Linden, & Geerlings, 2010; Janssen et al., 2000; Sinharay, Johnson, & Williamson, 2003), recently they have been extended to model the nesting of persons in groups while allowing measurement variant item parameters, resulting in multiple-group measurement models (De Boeck, 2008; De Jong & Steenkamp, 2010; De Jong, Steenkamp, & Fox, 2007; Fox, 2010; Fox & Verhagen, 2010; Frederickx, Tuerlinckx, De Boeck, & Magis, 2010; Verhagen & Fox, 2013a,b). Azevedo, Andrade, and Fox (2012) developed a generalized multiple group IRT model to handle response data from heterogenous groups with different latent means and variances, accounting for incomplete designs, and using more flexible population distributions when the normal distribution assumptions do not hold.

Let $j = 1, \dots, J$ denote the groups, then the probability of a correct response of person i in group j according to a multiple-group 1PL model is given by

$$P(Y_{ijk} = 1 \mid \theta_{ij}, \tilde{b}_{kj}) = \frac{e^{(\theta_{ij} - \tilde{b}_{kj})}}{1 + e^{(\theta_{ij} - \tilde{b}_{kj})}}, \tag{3}$$

with group-specific item (threshold) parameter $\tilde{b}_{kj}$.

A common model for the group-specific person parameters θ_{ij} in this model is a hierarchical model, in which the individual person parameters θ_{ij} are normally distributed around group means μ_j:

$$\theta_{ij} \sim N(\mu_j, \sigma_j). \tag{4}$$

As a hyperprior for μ_j, a normal $N(0, 1)$ distribution and as hyperprior for σ_j, an inverse gamma $IG(1, .1)$ distribution can be chosen, representing the expectation that the means will be on a standard normal scale. The next section will go deeper into the model for the group-specific item parameters $\tilde{b}_{kj}$.

2.4. Modeling the group-specific item parameters

When groups are considered a sample from a larger population, a multilevel or hierarchical structure can be assumed for the group-specific item parameters, to model measurement variance (De Jong et al., 2007; Fox, 2010; Fox & Verhagen, 2010; Verhagen & Fox, 2013a). For each item, group-specific deviations are assumed to be normally distributed with mean zero and variance $\sigma^2_{b_k}$. This variance component defines the variability in item functioning over groups in the population. If an item is invariant over groups, this variance component equals zero.

However, when there is interest in measurement invariance for specific groups, or if the number of groups is very small (creating difficulties in the estimation of the random item effect variances), a fixed instead of random group model is more useful. In fixed group models, all the group-specific parameters (group mean scores, group-specific item parameters) are estimated as separate parameters. The group-specific parameters in different groups are assumed to be independent and uninformative about each other, and there is no pooling of information across the groups. A possible prior distribution for the group means in this situation is a normal prior with a large variance parameter.

However, the group-specific item characteristics in the different groups are related when they refer to the same item. Items which are more difficult in one group will probably be among the more difficult items in other groups as well. Therefore, a multivariate normal model is imposed on the group-specific item characteristics, similar to the models for two groups proposed by De Boeck (2008) and Frederickx et al. (2010), in which the covariance matrix is used to model the correlation between item parameters from different groups. The group-specific item parameters are specified as:

$$\tilde{b}_{kj} = \mu_{b_j} + e_{kj},$$

where μ_{b_j} is the average item difficulty in group j, which is in our approach restricted to zero due to the choice of identification restrictions (Appendix A). The error e_{kj} represents the deviation of item difficulties from the general mean in group j. These deviations are assumed to be multivariate normally distributed with covariance $\boldsymbol{\Sigma}_b$.

This covariance matrix for the item difficulty parameters consists of the item parameter variance within each group ($\sigma^2_{b_j}$) on the diagonal, and the covariance of item parameters between each pair of groups $\sigma_{b_{jj'}}, j \neq j'$ on the off-diagonal:

$$\mathbf{e}_k \mid \boldsymbol{\Sigma}_b \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_b)$$

$$\boldsymbol{\Sigma}_b = \begin{bmatrix} \sigma^2_{b_1} & \sigma_{b_1 b_{\dots}} & \sigma_{b_1 b_J} \\ \sigma_{b_{\dots} b_1} & \sigma^2_{b_{\dots}} & \sigma_{b_{\dots} b_J} \\ \sigma_{b_J b_1} & \sigma_{b_J b_{\dots}} & \sigma^2_{b_J} \end{bmatrix}, \tag{5}$$

where $\mathbf{e}_k = (e_{k1}, \dots, e_{kJ})$. The variance of the item parameters over items ($\sigma^2_{b_j}$) can differ over groups, indicating that there is more variation in item difficulties in one group than in the other. The prior for the covariance structure $\boldsymbol{\Sigma}_b$ is of influence on our Bayes factor test, and we will therefore discuss the choice of this prior in Section 3.3.

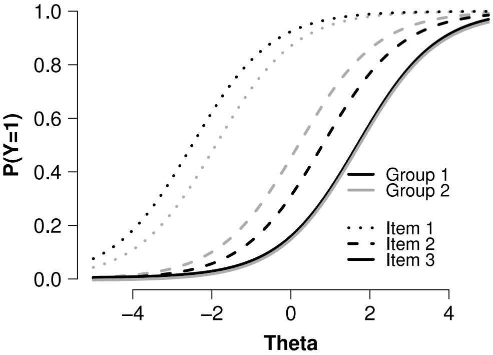

Fig. 2. An illustration of the Bayesian IRT model for fixed groups. The lines represent the item characteristic curves associated with the item difficulties for group 1 (black) and group 2 (gray) of three different items.

As separate item difficulty parameters are estimated for each of the selected groups, measurement invariance can be tested by comparing group-specific difficulty parameters directly. To facilitate this comparison, a difference parameter $d_{k_{jj'}}$ can be defined as the difference between two group-specific item parameters in group j and j' ($j < j'$):

$$d_{k_{jj'}} = b_{kj} - b_{kj'}, \quad j < j'.$$

In the following, the description of invariance tests will be limited to a situation with two groups, which results in a single difference parameter d_k per item K. The resulting prior distribution for this difference is described in Section 3.3. The result can be extended to more groups, however, as is described in Appendix E.

The structure of the item parameters is illustrated in Fig. 2. Presented are the item characteristic curves of three hypothetical items for two groups, indicating how the probability of endorsing an item (answering the item correctly) changes as a function of the ability parameter. The gray and black lines represent the group-specific item characteristic curves with difficulty parameters $\tilde{b}_{kj}$, which are random but correlated deviations from the average difficulty parameter (set to zero). The item characteristic curves differ due to between-item and between-group differences in item difficulty. Item 3 in the figure is invariant, as the groups have equal difficulty parameters for this item, and the item characteristic curves overlap. A test of the difference between the difficulty parameters would support the null hypothesis of invariance. The item difficulty of Item 1 in group 1 is higher than the difficulty in group two, while the item difficulty of item 2 is lower in group 1 than in group 2, indicating measurement variance.

To make sure the parameters in the model are identified, the group-specific item parameters are restricted to sum to zero within each group (for each group j, $\sum_k \tilde{b}_{kj} = 0$). The underlying assumption is that for each group, the test as a whole has the same difficulty level, where individual items can deviate from this average within groups. An advantage is that no specific parameters need to be fixed, implicating that no invariant items or reference groups have to be specified beforehand. This makes the anchoring more robust in case no information is known about anchor items beforehand. The absence of individually fixed parameters also makes it easy to include explanatory variables to account for group differences in latent means and item parameters. It is possible, however, to replace this restriction with an anchor item restriction in the presented models when anchor items are known. More information about the identification of multiple-group IRT models can be found in Appendix A. The WinBugs specification for the model in this section can be found in Appendix B.

3. Testing for measurement invariance

An item is measurement invariant when persons from each group with the same true value on the measured construct have the same probability of endorsing that item. When testing an item for measurement invariance in two groups, the null hypothesis of invariance corresponds to equality of the item parameter in both groups ($H_0 : d_k = 0$), while the alternative hypothesis states that the item parameters in the two groups differ ($H_1 : d_k \neq 0$).

Many methods have been developed to assess whether items within a test exhibit measurement invariance, or, equivalently, whether the items are free of differential item functioning (dif). Well-known methods are based on nonparametric analysis, linear regression, factor analysis and item response theory (IRT) (see for an overview: Millsap, 2011). The focus here will be on measurement invariance for IRT models. One widely used parametric method to test for measurement invariance is the likelihood ratio test for measurement invariance (Thissen, Steinberg, & Wainer, 1993), based on maximum likelihood estimation methods (f.e. EM algorithm, Bock & Aitkin, 1981). To establish which item parameters are invariant, multiple models with varying numbers of items restricted to invariance are compared with respect to their goodness of fit. Recently, invariance tests based on the Wald test (Lord, 1980) to test the equality of item parameters were developed (Woods et al., 2012) by improved estimates of the covariance matrix (Cai, 2008) and extension of the test to allow for comparison between more than two groups at a time (Kim, Cohen, & Park, 1995). These tests are implemented in the software package IRTpro (Cai et al., 2011). A very different method to test measurement invariance was recently implemented in Mplus (Muthén & Muthén, 2012), introduced as the alignment method (Muthén & Asparouhov, 2013b). After estimating a solution in which all parameters are allowed to vary over groups, a rotation (similar to rotations in exploratory factor analysis) is applied aimed at either small or large differences between groups in item parameters. Testing for invariance proceeds by examining significant differences between individual item parameters.

There are two main limitations of these current invariance tests. First, It is not possible to gather evidence in favor of the invariance hypothesis ($H_0 : d_k = 0$). Second, some invariant (anchor) items need to be identified beforehand (except for the alignment method).

Bayesian hypothesis testing using the Bayes factor has many advantages in measurement invariance testing. Assume a situation in which a sample of American and a sample of Chinese students are compared on mathematics ability, measured by twenty mathematics items. Beforehand, there is no information about the invariance of any of these items. To investigate measurement invariance of this instrument, the ideal invariance test would evaluate the hypothesis of invariance directly, for all items simultaneously, and without the need to specify invariant items in advance. The combination of these characteristics can only be achieved through Bayesian hypothesis testing using identification restrictions on the item parameters.

After outlining the advantages of the Bayes factor test, a new Bayes factor test for group-specific differences will be introduced, which can be used when selected groups are of specific interest. Informed prior choices for the variance components will be discussed in the last part of this section.

3.1. Advantages of Bayesian tests for measurement invariance

3.1.1. Evidence favoring invariance

Current measurement invariance tests for IRT models are based on frequentist statistical theory. A sampling distribution is assumed for the test statistic of interest (e.g. Wald, χ^2) under the null hypothesis, representing the distribution of this test statistic if the item parameters were invariant. The null hypothesis is rejected when it is very unlikely (for example, $p < .05$) that the observed value of the test statistic or a more extreme value is found in a sample from a population in which the null hypothesis holds. In case of measurement invariance testing, however, the evidence *in favor* of the invariance hypothesis is of main interest. Furthermore, it has been shown that evaluating evidence only with regard to the null hypothesis without taking evidence about the alternative hypothesis into account leads to results which overstate the evidence against the null hypothesis (Rouder, Speckman, Sun, Morey, & Iverson, 2009; Sellke, Bayarri, & Berger, 2001; Wagenmakers et al., in press) especially in low powered studies (Wagenmakers et al., 2014).

Using a Bayesian approach to hypothesis testing enables researchers to investigate invariance directly. The amount of evidence in favor of invariance is quantified by a Bayes factor (Jeffreys, 1961; Kass & Raftery, 1995). Instead of focusing on rejecting one hypothesis, the Bayes factor evaluates evidence for both the null and the alternative hypothesis given the data. The result of the test can indicate convincing support for either the null or alternative hypothesis, or evidence can be inconclusive about which hypothesis is preferred.

In tests of invariance the null hypothesis can be specified as a point hypothesis (the difference between two item parameters is zero), and the alternative hypothesis as an area (the difference between two item parameters is not zero). The marginal likelihood of the alternative hypothesis given the data can be defined as the average likelihood over all plausible values of the alternative hypothesis, weighted by the prior probability assigned to these values. This average likelihood equals the integral of the likelihood function weighted by the prior density function over the parameter space contained by the hypothesis. The ratio of the marginal likelihoods for both hypotheses results in the Bayes factor. For inference about the difference d_k based on data Y, with null hypothesis $H_0 : d_k = 0$, and alternative hypothesis $H_1 : d_k \neq 0$, the Bayes factor can be expressed as:

$$BF_{01} = \frac{P(Y \mid H_0)}{P(Y \mid H_1)} = \frac{P(Y \mid d_k = 0)}{\int P(Y \mid d_k) p_1(d_k) dd_k}, \qquad (6)$$

where $p_1(d_k)$ is the prior distribution under the alternative hypothesis. The prior distribution for d_k will be further discussed in Section 3.3. It follows that the Bayes factor gives a direct measure of the relative evidence in favor of the null and the alternative hypothesis.

3.1.2. Avoiding anchor items

Each measurement model used for invariance testing which contains both latent group means and/or variances and group-specific item parameters has an identification problem. There are several ways to resolve this problem, which are described in more detail in Appendix A. Most current measurement invariance tests need at least one invariant "anchor" item to link the scales of the groups under comparison. Unless prior knowledge exists about the invariance of certain items, one has to resort to empirical anchor selection methods, which can be tedious, especially if items have to be invariant over a large number of groups. Furthermore, the accuracy of the selection will influence the results of the invariance test. Langer (2008) introduced a two-step procedure in which first

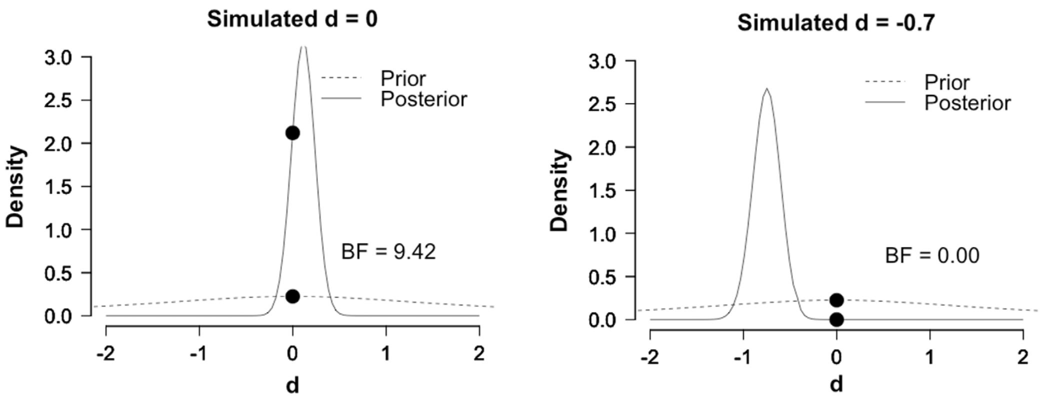

Fig. 3. Illustration of the Bayes factor test for the difference d_k between the item parameters of two groups for item k. The dots indicate the density of the prior (dotted line) and posterior (solid line) at the null hypothesis $d_k = 0$, for an item with no difference $d_k = 0$ and for an item with a large difference $d_k = -.7$ between the item parameters of two groups.

group means are estimated under a fully invariant model, and then invariance tests are performed with group means fixed to these values. This can lead to biased estimates, however, when there is a substantial amount of DIF (e.g. Woods et al., 2012).

Replacing invariance restrictions with restrictions on the sum of the difficulty parameters eliminates the need to fix model parameters. The underlying assumption is that for each group, the test as a whole has the same average level of difficulty. An advantage is that no model parameters need to be fixed, implicating that no invariant items or reference groups have to be specified beforehand, and between-group differences in item characteristics are allowed. This makes the test anchoring more robust in case there are no anchor items known beforehand (Appendix A), which makes the restriction particularly useful for exploratory analyses. The result of this different restriction is that measurement invariance can be evaluated for all K items simultaneously.

3.2. Bayes factor for measurement invariance

Bayes factors are not always easy to compute. In case of nested models, however, it can be shown that the Bayes factor in favor of the null hypothesis reduces to the ratio of the density (or density region) of the null hypothesis under the posterior and prior distribution of the most complex model, the Savage–Dickey density ratio (see e.g. Dickey, 1971; Verdinelli & Wasserman, 1995). The Savage–Dickey density ratio can be used to construct a test for measurement invariance. The null hypothesis of invariance of an item for any two groups in the fixed groups model can be defined as the difference between the item parameters of two groups ($d_k = b_{k1} - b_{k2}$) being equal to zero. To evaluate the relative support for the null hypothesis over the alternative hypothesis, the Bayes factor reduces to the ratio of the density of the null hypothesis ($d_k = 0$) under the posterior $p_1(d_k \mid H_1, Y)$ and prior $p_1(d_k \mid H_1)$ distribution of the difference d_k under the alternative hypothesis H_1:

$$BF_{01} = \frac{p_1(d_k = 0 \mid H_1, Y)}{p_1(d_k = 0 \mid H_1)}. \quad (7)$$

The invariance of all K items can be evaluated simultaneously, as the Bayes factors are computed based on the marginal prior and posterior distributions of the differences in item parameters. The prior distribution $p_1(d_k \mid H_1)$ is discussed further in 3.3.

Within an MCMC sampling scheme, there are several ways to compute these Bayes factors for nested models. One relatively easy computation method is to sample from the posterior distribution for d_k under the alternative hypothesis, for example using WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000). Subsequently the density at the null hypothesis under both models can be computed using for example R (Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010; Wetzels, Raaijmakers, Jakab, & Wagenmakers, 2009) (Appendix C).

The interpretation of the Bayes factor tests for measurement invariance is straightforward. The Bayes factor indicates how much more likely the data are given the null hypothesis than given the alternative hypothesis. The categorization proposed by Jeffreys (1961) provides a guideline to decide whether there is substantial evidence for either hypothesis. Following this categorization, a Bayes factor (BF_{01}) larger than three, implying that the data are three times more likely under H_0 than under H_1, is considered substantial support for the null hypothesis H_0, while a Bayes factor larger than 10 is considered strong support for H_0. A Bayes factor smaller than .33 is considered as substantial support for the alternative hypothesis H_1 ($BF_{10} = 1/BF_{01}$ is larger than 3) and a Bayes factor smaller than .01 as strong support for H_1. This categorization is arbitrary, however, and can be replaced by more or less conservative criteria based on what is considered "strong evidence" in light of the hypothesis under investigation.

Fig. 3 gives an illustration of the way the Savage–Dickey density ratio works. In the figure on the left (for the situation in which data are simulated for $d_k = 0$), the density of the posterior distribution (solid line) at the null hypothesis $d_k = 0$ is higher than the density of the prior distribution (dotted line) at H_0. The Bayes factor is 9.42, indicating substantially more evidence for H_0 than for H_1. In the figure on the right, where the data have been simulated with $d_k = .7$, the prior density at the null hypothesis is higher than the posterior density at that point. The Bayes factor less than .001 indicates strong support for the alternative hypothesis H_1.

3.3. Choosing priors

The ratio of the density of the null hypothesis under the posterior and prior distributions, and therefore the result of the Bayes factor test, depends on the priors chosen for the parameters under evaluation. Priors can be chosen, however, to reflect reasonable assumptions about the parameter values.

The prior distribution for the difference in item parameters d_k results from the prior distributions of the separate item parameters. The prior for the separate group-specific item parameters is set to a normal distribution, with mean $\mu_{b_j} = 0$ (as is part of the restriction discussed in Appendix A) and a covariance matrix $\boldsymbol{\Sigma}_b$. Therefore, the prior specified for $\boldsymbol{\Sigma}_b$ is important for the results of the Bayes factor. Two different priors for this covariance matrix will be compared.

The first prior under consideration takes the identity matrix, which has diagonal ones, as the prior covariance matrix. As the off-diagonal zero's represent a covariance of zero between the

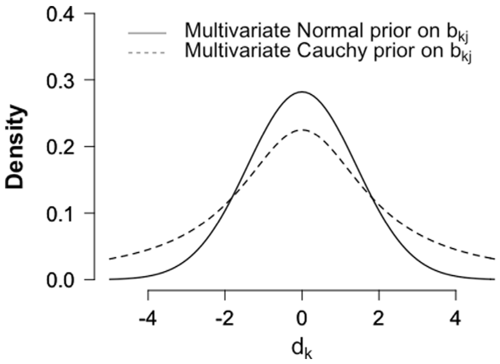

Fig. 4. Illustration of the marginal prior distributions for the difference between item parameters using a standard multivariate normal prior or a multivariate Cauchy prior on the item parameters.

item parameters, the marginal prior distributions of the two item parameters are independent standard normal distributions. Hence, the distribution of the difference between the item parameters is a univariate normal distribution with a mean of zero and a variance of two: $N(0, 2)$. This prior represents the assumption that the item parameters will be approximately standard normally distributed, which corresponds to the usual range of item parameters observed in 1PL models.

A conjugate prior for the covariance of a multivariate normal distribution is an Inverse Wishart prior $IW(S, J)$. This results in a multivariate Cauchy prior on the item parameters $\tilde{b}_{kj}$, a multivariate normal distribution with mean zero and an Inverse Wishart prior on the variance. When an S matrix with ones on the diagonal and zeros elsewhere is chosen, the resulting distribution of d_k is a non-standardized univariate t distribution with 1 degree of freedom and a scale parameter of 2. This result can be extended to comparisons between more than two groups (see Appendix E).

Fig. 4 shows the distribution of the difference between item parameters for two groups with this multivariate Cauchy prior on the group-specific item parameters, and how it compares to the multivariate normal prior on the group-specific item parameters. Although both distributions cover a similar range of plausible values for d_k, the multivariate Cauchy distribution has wider tails and lower prior density around zero. The prior is less informative about the difference, as the density is more equally spread. Therefore, the Bayes factor will be more likely to favor invariance under the multivariate Cauchy prior than under the multivariate normal prior on the item parameters.

4. Simulation study: Evaluation of the Bayes factor test for item parameter differences

A simulation study was performed to evaluate the Bayes factor test for item parameter differences. Two different priors were evaluated for the variance of the item parameters (see 3.3). A comparison is made with invariance tests based on the Wald test (Lord, 1980) to test the equality of item parameters (Woods et al., 2012) as implemented in IRTPRO (Cai et al., 2011). In this test, all other items are considered anchor items while testing each item for invariance. The Bayes factors were computed using a procedure in R based on WinBUGS output (Wetzels, Grasman, & Wagenmakers, 2010; Wetzels et al., 2009) (Appendix C). The Bayes factors (BF_{01}) compare the null hypothesis of invariance ($H_0 : d_k = 0$) to the alternative hypothesis that there is a difference between the item parameters of two groups ($H_1 : d_k \neq 0$).

Data were generated consistent with two sets of assumptions: for each group j, the sum of the item thresholds $\sum_k b_{kj}$ equaled zero, the mean for the reference group μ_{θ_j} equaled zero and two (anchor) items were invariant. Data for two groups both consisting of 250, 500 or 750 subjects answering ten items were generated, with the ten items containing five pairs of items with an increasing difference d_k between the item parameters of the two groups (0, .1, .3, .5 and .7). The combined measurement invariance test results of the analysis of 50 simulated data sets are presented in Table 1. The results for estimation accuracy are in Appendix D.

The right-hand part of Table 1 shows results of the Bayes factor test using a Cauchy prior on the item difficulties, and the middle columns show results of the Bayes factor test using a Normal prior (see Section 3.3). Although some prefer a higher cut-off (e.g. 10) for substantial evidence, or a continuous interpretation of the Bayes factor, we chose a minimal cut-off of 3 for the invariance and .33 for the non-invariance hypothesis to indicate sufficient evidence to support the hypotheses. In the left-hand part of the table, results acquired with the maximum likelihood based EM algorithm in IRTPRO (Cai et al., 2011) are shown, using a Wald test to evaluate each item for invariance, with all other items as anchors. Cut-off p-values (α) for rejection of the null hypothesis of both .01 and .05 are shown.

Of main interest is whether the Bayes factor test will identify the *invariant items* ($d_k = 0$) by high support in favor of the null hypothesis. Using a multivariate Cauchy prior on the item difficulties, for 91%–97% of the invariant items there was substantial evidence for the null hypothesis of invariance, as indicated by a Bayes factor larger than three. None of the invariant items were incorrectly identified as non-invariant items (indicated by a Bayes factor lower than .33), leaving three to nine percent of the items undecided (not enough evidence for either hypothesis). As expected, the more informative multivariate standard normal prior on the item difficulties rendered the support for invariance less convincing, especially for the smaller group sizes. Under this prior, 78%–91% of the invariant items were identified as invariant. Also, one percent of the invariant items was wrongly indicated as non-invariant using this prior, leaving 9%–22% of the invariant items with no clear support for invariance or non-invariance. The Wald test rejected the null hypothesis incorrectly for one to two percent of the invariant items at $\alpha = .05$, and for none of the items at $\alpha = .01$. While the performance of the Wald test is similar to the performance of the Bayes factor with regard to falsely reporting non-invariance (or type I errors), the Wald test misses the ability to evaluate the strength of evidence in favor of invariance.

Overall, the percentage of Wald tests which reject the null hypothesis of invariance corresponds closely to the percentage of Bayes factor tests (based on the Cauchy prior) confirming the hypothesis of variance in item parameters (given that a Bayes factor of 3 is considered strong enough evidence to support a hypothesis). The Wald tests show an increasing percentage of null hypothesis rejections when the difference between item parameters grows, and near perfect detection of non-invariant items with a difference in item parameters of .7, especially with larger sample sizes. The Bayes factors, however, give information about both hypotheses by being more often indecisive for small differences in item parameters ($d_k = .1$ or $d_k = .3$) and clearly favoring measurement variance for large differences in item parameters ($d_k = .5$ or $d_k = .7$). Comparing the two priors, the Bayes factor based on the Normal prior is more conservative in its support for the invariance hypothesis, as the prior has higher density at the null hypothesis.

As the Wald test can only reject invariance but not support variance, and the Bayes factor is the ratio of the amount of support for either hypothesis, the results should be compared with caution. Both tests perform very well in identifying large differences ($d_k \geq .5$) between item parameters, especially with larger sample sizes. The Bayes factors are more conservative when invariance decisions would be made based on substantial evidence for H_0, but less conservative when decisions would be made based on whether

Table 1
Results for the Bayes factor test for item parameter differences with two different priors and the Wald test for five pairs of items with increasing amounts of DIF between two groups, over 50 replicated data sets.

	BF_{01} Cauchy prior		BF_{01} Normal prior		Wald test	
750 subjects per group						
d_k	$\%BF_{01} > 3$	$\%BF_{01} < .33$	$\%BF_{01} > 3$	$\%BF_{01} < .33$	$\%p < .05$	$\%p < .01$
0.00	0.94	0.00	0.91	0.01	0.01	0.00
0.10	0.81	0.01	0.75	0.06	0.08	0.01
0.30	0.23	0.38	0.23	0.44	0.61	0.31
0.50	0.00	0.94	0.00	0.91	0.99	0.94
0.70	0.00	1.00	0.00	1.00	1.00	0.99
500 subjects per group						
d	$\%BF_{01} > 3$	$\%BF_{01} < .33$	$\%BF_{01} > 3$	$\%BF_{01} < .33$	$\%p < .05$	$\%p < .01$
0.00	0.91	0.00	0.86	0.01	0.02	0.00
0.10	0.90	0.01	0.74	0.01	0.05	0.00
0.30	0.33	0.25	0.22	0.31	0.49	0.19
0.50	0.02	0.71	0.08	0.73	0.85	0.67
0.70	0.00	0.95	0.00	0.96	0.98	0.95
250 subjects per group						
d	$\%BF_{01} > 3$	$\%BF_{01} < .33$	$\%BF_{01} > 3$	$\%BF_{01} < .33$	$\%p < .05$	$\%p < .01$
0.00	0.97	0.00	0.78	0.01	0.01	0.00
0.10	0.81	0.04	0.74	0.03	0.11	0.02
0.30	0.42	0.17	0.39	0.19	0.38	0.18
0.50	0.14	0.68	0.12	0.43	0.85	0.71
0.70	0.02	0.93	0.04	0.88	0.99	0.95

or not there is substantial evidence for H_1, especially when the α for the rejection of the null hypothesis would be .05. For the smaller differences ($d_k = .1$, $d_k = .3$), the Bayes factor is undecided for a percentage of items, which is desirable as it represents the uncertainty of the situation. Overall, the tests produced practically similar results, given that the assumptions were met for all models. An advantage of the Bayes factor is that it gives more detailed information about the support for both hypotheses. When the goal is to identify which items are anchor items, the possibility to evaluate the evidence in favor the null hypothesis is exactly what is desired.

5. Analyzing geometry items for males and females (CBASE)

To illustrate the use of the multiple-group IRT model and Bayes factor tests for measurement invariance in real test situations, geometry items from the College Basic Academic Subjects Examination (CBASE) for males and females are analyzed (Flowers, Osterlind, Pascarella, & Pierson, 2001; Millsap, 2011; Osterlind, Robinson, & Nickens, 1997). CBASE is an exam intended for students enrolled in college, assessing knowledge and skills in mathematics, English, science and social studies. The analysis will focus on measurement invariance for 11 items from the geometry subtest of the mathematics test, comparing females ($N = 4452$) and males ($N = 1034$). Bayesian IRT models for fixed groups will be estimated and the corresponding Bayes factors to test measurement invariance will be computed. As the Cauchy prior showed better results in the simulation study, only this prior was used to analyze these example data. The results will be compared to results from a traditional analysis based on maximum likelihood estimates, in which the scales are linked with anchor items instead of by equal average thresholds in both groups.

After 5000 iterations, with a burn in of 500 iterations, convergence of the MCMC chains was reached, as all lag 50 autocorrelations were below .1 and all Geweke Z statistics (Cowles & Carlin, 1996) below 2. Table 2 shows the results for this model.

The Bayes factor tests convincingly identify item 6 and 7 as non-invariant items, whereas for item 2, 3, 4, 5 and 9 there is substantial evidence that they are invariant over groups. Item 7 is easier for male students, while item 6 is easier for female students.

The same CBASE data set was analyzed with standard maximum likelihood based procedures (EM algorithm, Bock & Aitkin, 1981) with anchor item and reference group restrictions as implemented in IRTPRO (Cai et al., 2011). First, all items were tested for invariance, using all other items as anchors. Column 8 in Table 2 shows the results of this Wald test. Even though the restrictions on which the parameters and parameter differences for these tests are based were different from the restrictions for the Bayes factor test, the results are remarkably similar. For item 6 and 7, invariance would be rejected based on p-values $< .01$, and additionally for items 1 and 11 based on a p-value $< .05$, while for items 3, 4, 5 and 9 the p-values are highest.

Next, the items for which invariance was clearly not rejected were used as an anchor set to estimate item parameters for all items. Comparing these parameter estimates with the estimates from the Bayesian IRT model (Table 2), it is clear that the mean and variance of the scales are different, and as a result, the maximum likelihood item parameter estimates are higher and less spread out. This is a direct result of the identification restrictions: the sum of the item parameters was set to zero in the Bayesian IRT model, while the mean of the θ scale for males was set to zero in the maximum likelihood method.[1]

The last columns of the table show the parameters rescaled to the scale of the Bayesian IRT estimates, by subtracting within each group the mean difficulty from each group-specific item parameter, and multiplying by the ratio of the standard deviations of θ in that group. When compared, the rescaled parameters are almost equal to the Bayesian IRT model estimates. This shows that the large amount of data dominates the posterior information and the priors are not influential. The rescaled anchor items are not exactly equal for males and females anymore, however, and this is a direct consequence of the point at which the scales are linked, reflecting either a set of anchor items or an equal overall difficulty of the test. As there are many items and a relatively large amount of invariant items in this example, the invariance tests give exactly the same result.

[1] Another difference is that the discrimination parameter is set to 1 in the Bayesian IRT models but is estimated in IRTPRO, resulting in the different variances.

Table 2
Bayesian and ML Item parameter estimates and results for the Bayes factor test and Wald test for item parameter differences (CBASE example).

Item	Bayesian IRT				Maximum Likelihood					Rescaled ML		
	$\hat{b}_k$ M	$\hat{b}_k$ F	DIF	BF_{01}	$\hat{b}_k$ M	$\hat{b}_k$ F	DIF	χ^2(df)	p	$\hat{b}_k$ M	$\hat{b}_k F$	DIF
1	−0.27	−0.03	−0.24	0.43	−1.17	−0.95	−0.22	5.20(1)	0.02	−0.28	−0.04	−0.24
2	0.78	0.67	0.11	7.25	−0.45	−0.45	0	2.30(1)	0.13	0.72	0.67	0.05
3	−0.66	−0.77	0.11	9.16	−1.46	−1.46	0	1.20(1)	0.28	−0.68	−0.76	0.09
4	0.68	0.63	0.05	18.04	−0.48	−0.48	0	0.70(1)	0.40	0.68	0.63	0.05
5	−1.09	−1.25	0.16	4.82	−1.79	−1.79	0	1.90(1)	0.17	−1.14	−1.23	0.10
6	0.46	0.16	0.30	0.07	−0.63	−0.81	0.18	12.20(1)	0.00	0.47	0.16	0.31
7	0.85	1.21	−0.36	0.02	−0.36	−0.07	−0.29	13.50(1)	0.00	0.85	1.22	−0.37
8	1.11	1.28	−0.17	2.17	−0.05	−0.05	0	2.40(1)	0.12	1.28	1.24	0.04
9	−0.61	−0.68	0.07	14.48	−1.40	−1.40	0	0.50(1)	0.47	−0.59	−0.68	0.08
10	−0.87	−1.06	0.19	2.52	−1.66	−1.66	0	3.20(1)	0.07	−0.96	−1.05	0.09
11	−0.38	−0.17	−0.21	1.10	−1.24	−1.04	−0.20	3.90(1)	0.05	−0.37	−0.17	−0.21
μ_{θ_j}	1.36	0.62			0.00	−0.49						
σ_{θ_j}	1.47	1.28			1.06	0.90						

There are situations, however, in which different linkage rules can lead to different test results. In this case, in which there are often many non-invariant items, the equal overall difficulty restriction creates the possibility to paint an overall picture of differences in item parameters between groups independent of the chosen anchor items, and the possibility of including explanatory information directly into the model (see also Verhagen & Fox, 2013a). When the aim is to identify and use anchor items, the average difficulty restricted Bayesian IRT models can be used as a base to start exploring which items are invariant. Parameters can then be restricted to invariance in a second step. In this example, items 4 and 9 are clearly indicated to be invariant, and could be used as anchor items in a second estimation round for Bayesian IRT models, using an anchor item restriction instead of or in addition to the equal average difficulty restriction.

6. Discussion

A Bayes factor was developed to test for measurement invariance in IRT models. Using a Bayesian multiple-group IRT model for fixed groups, in which the group-specific item parameters are assumed to be multivariate normally distributed, measurement invariance is tested by evaluating differences in group-specific difficulty parameters.

The Bayes factor test for measurement invariance was evaluated and compared to the Wald test as implemented in IRTPRO. The simulation study showed results for the Bayes factor similar to the results from the Wald tests (assuming a Bayes factor of 3 to be substantial evidence to support a hypothesis). The results should be compared with some caution, however, as the Wald test advises whether or not to reject the null hypothesis, while the Bayes factor evaluates the relative evidence for the null and alternative hypotheses. The Bayes factor is more conservative in indicating invariance than the Wald test when decisions would be made based on substantial evidence for H_0, but less conservative when decisions would be made based on whether or not there is substantial evidence for H_1. Especially when the goal is to identify anchor items, the possibility of evaluating evidence in favor of invariance is desired. Another advantage of the Bayes factor is that all parameters can be tested for invariance simultaneously, which makes the test especially useful for exploratory purposes.

The CBASE example illustrated the use of the Bayes factor tests for invariance. The Bayesian IRT models produced approximately the same Bayes factor test results as the maximum likelihood based estimation and Wald tests. This showed that the amount of data dominated the posterior information and the priors are not very influential in the estimation process. As the scales are linked at a different point, choice for one or the other linkage rule can result in different parameter estimates and therefore can point to different items as non-invariant items. However, in situations with many (invariant) items and groups, the differences are often minor. Both linkage rules can be used in the Bayesian IRT models. It is up to the researcher to determine which rule to choose, based on theoretical as well as practical arguments. When interest is in identifying invariant items, a two-step procedure can be implemented. First, the equal threshold restriction in combination with the Bayes factor test can be used to indicate which items are most likely to be invariant. Then, in a second step, the anchor item restriction can be used to estimate the item parameters in the final model.

Another Bayesian two-step procedure was proposed by Muthén and Asparouhov (2013a) as implemented in Mplus (Muthén & Muthén, 2012). Following an initial Bayesian estimation using very broad priors on free parameters and very narrow priors on fixed parameters, models with "approximate" measurement invariance are estimated. They do not use a Bayes factor to identify variance in parameters over groups, however, but a significant ratio of the difference between two individual parameters and its standard error, which is similar to null hypothesis significance testing.

Future research could extend the framework further, with for example multi-dimensional IRT models or to measurement instruments with a mixed number of answer categories. Issues which have been encountered while investigating these models, like the prior sensitivity of the Bayes factors and the effect of linkage rules under different conditions (the size of differences between parameters, the number of non-invariant items, the amount of groups) could be investigated in more detail.

This paper showed that the Bayes factor is a valid alternative to current measurement invariance tests. The advantages of evidence in favor of the null, the possibility to test all items simultaneously and the average threshold restriction makes the Bayes factor especially useful for exploratory research in the absence of knowledge about anchor items, and in cases where explanatory information can be included to explain differences in item parameters between groups.

Appendix A. Identification of multiple-group IRT models

Each measurement model containing both latent person parameters and item threshold parameters has an identification problem. Several combinations of item parameters and latent variable values result in identical likelihood values, complicating parameter estimation. In single group settings, this identification problem is generally solved by fixing the latent variable to have a mean of zero. As the mean of the scale is arbitrary, this restriction has no implications for the interpretation of the model.

In a multiple group setting, however, this identification problem exists within each group. In addition, the scores of the different groups have to be estimated on the same scale. There are several ways in which the multiple-group 1PNO IRT model can be identified. First, the scale has to be identified for one group by fixing at least one group-specific parameter. This can either be the group mean ($\mu_{\theta_1} = 0$) of the person parameters, the sum of the item thresholds for that group ($\sum_k \tilde{b}_{k1} = 0$), or one (or more) of the group-specific threshold parameters (e.g. $\tilde{b}_{k1} = 0$). Once the scale for one group has been identified, the scales of the other groups can be linked as well as identified by defining either at least one group-specific item parameter (e.g. $\tilde{b}_{k1} = \tilde{b}_{kj}$), the sum of the item parameters within the group ($\sum_k \tilde{b}_{k1} = \sum_k \tilde{b}_{kj}$), or the person parameter mean ($\mu_{\theta_1} = \mu_{\theta_j}$) to be equal to that of the first group.

Traditional methods to estimate multiple-group IRT models and test for measurement invariance (e.g. likelihood ratio test (Thissen et al., 1993) or Wald test (Lord, 1980; Woods et al., 2012)), based on maximum likelihood estimates of the item parameters, are usually identified based on a fixed person parameter mean μ_θ for a reference group and at least one anchor item with equal item parameters in all groups. Although easily implemented in an ML estimation procedure, there are some disadvantages to these identification restrictions. Restricting the mean and variance for one group complicates the modeling of person parameters for the measured construct, like explanatory covariates or multilevel (longitudinal) structures. Unless prior knowledge exists about the invariance of certain items, one has to resort to empirical anchor selection methods, which can be tedious, especially if items have to be invariant over a large number of groups. Furthermore, the accuracy of the selection will influence the results of the invariance test. Langer (2008) introduced a two-step procedure in which first group means are estimated under a fully invariant model, and then invariance tests are performed with group means fixed to these values. This can lead to biased estimates, however, when there is a substantial amount of DIF (e.g. Woods et al., 2012). In addition, if the wrong items are chosen as anchor items, or if none of the items is invariant, this causes bias in the estimated latent scores and in the estimated group differences.

In the Bayesian random item parameter framework, restricting the mean of the threshold parameters to zero within all groups is a natural choice, which reflects the assumption of equal test difficulty across groups. This leaves the variance components free to be estimated in a very flexible modeling framework. The restrictive assumption is spread out over all the item parameters, which leads to more robust anchoring in case there are no anchor items known beforehand. In case the assumption is wrong, and there is for example only one item which is not measurement invariant, this leads to less bias than when an item is falsely restricted to be an anchor item. When there is only one item variant, that item will show a deviation in one direction in one or more groups, while all the other (invariant) items are forced to make a small deviation in the other direction in these groups to preserve the overall constraint. The deviations of the invariant items become smaller when the number of items increases and will always be relatively small in comparison with the real variant item. Once the real variant item is identified, it is possible to anchor the other items and estimate the item parameters accurately.

Appendix B. Model specification in WinBUGS

Fixed multiple-group IRT models

This section will present the WinBUGS (Lunn et al., 2000) code of the fixed manifest groups model for a data set Y with J groups j, K items k, and N persons i, stacked in such a way that *njl*[1] is the first person for group j and *njh*[*J*] is the last person for group j.

1. Basic model

```
for (j in 1:J){
  for (i in njl[j]:njh[j]){
      for (k in 1:K){
          logit(p[i,k]) <- theta[i]-Rbeta[k,j]
          Y[i,k] ~ dbern(p[i,k])
          }
      theta[i] ~ dnorm(mut[j],prec[j])
}}
```

2. Priors for group means

```
for (j in 1:J){
    mut[j] ~ dnorm(0,1)
    prec[j] ~ dgamma(1,.1)
    sigmat[j] <- 1/prec[j]
}
```

3. Item parameters

```
for (k in 1:K){
   beta[k,1:J] ~ dmnorm(mu[],Prec[,])
   priorbeta[k,1:J] ~ dmnorm(mu[],Precprior[,])
}
```

4. Rescale item parameters to the accommodate the restriction that for each group *j*, $\sum \tilde{b}_{kj} = 0$.

```
for (j in 1:J){
    meanb[j] <- mean(beta[1:K,j])
for (k in 1:K){
    Rbeta[k,j] <- beta[k,j] - meanb[j]
}}
```

5. Model DIF for Bayes Factor, 2 groups

```
for (k in 1:K){
    dif12[k] <- Rbeta[k,1]- Rbeta[k,2]
}
```

6. Priors for the item parameters: multivariate Cauchy prior (Section 3.3), R= $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

```
for (j in 1:J){mu[j] <- 0}
Prec[1:J,1:J] ~ dwish(R[1:J,1:J],J)
Sigma[1:J,1:J] <- inverse(Prec[1:J,1:J])
```

7. Priors for the item parameters: multivariate Normal prior (see Section 3.3), R= $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

```
for (j in 1:J){mu[j] <- 0}
Sigma[1:J,1:J] <-  R[1:J,1:J]
```

Appendix C. Bayes factor computation based on WinBUGS output in R

Bayes factor test for item parameter differences

The Bayes factors comparing nested models with regard to the difference in item parameters for all items simultaneously can be specified in R based on the index (index) and coda (coda) files of WinBUGS (Lunn et al., 2000) output with *XG* iterations and *BURN* as the number of burn in iterations. The sampled prior and posterior values of the differences in item parameters are specified in the variable "difchains", after which the density at $d_k = 0$ under both distributions under the logspline approximation of the density is computed similar to Wagenmakers et al. (2010).

```
# Define K (the number of items) , XG (the number of iterations)
# and BURN (the number of burn in samples)

K = 10
XG = 5000
BURN = 1000

# index: matrix with indices which indicate the start and end of the chains
# for the parameters in the coda matrix (Read WinBUGS Index file)
# coda =the matrix with the MCMC chains (Read WinBUGS coda file)

index <- read.table(file=tindex, sep="␣",
  row.names=2)
coda <- read.table(file=tcoda , sep="")

# Read the MCMC chains for the difference in item parameters

lo <- which(rownames(index) == "dif12[1]")
hi <- lo+K-1
it <- XG-BURN
difchains <- matrix(0,it,K)
difchains[,1:K] <- matrix(coda[ind[lo,2]:
  ind[hi,3],2],it,)

# set the  value of the prior density at H0: Cauchy prior on the item parameters
if (whichprior == "CP") {
prior <-(1/(sqrt(2)*pi)) }

# value of the prior density at H0: multivariate normal prior on the item parameters
if (whichprior == "NP") {
prior <- dnorm(0,0,sqrt(2)) }

# Use polspline to calculate the density of the posterior at H0
# and compute the Bayes factor (see also paragraph 3.2)

posterior <- matrix(0,K,1)
BF01 <- matrix(0,K,1)
for (k in 1:K){
 fit.posterior <- logspline(difchains[,k])
 posterior[k] <- dlogspline(0,fit.posterior)
 BF01[k] <-posterior[k]/prior
}
```

Appendix D. Simulation results: estimation accuracy

See Table 3.

Appendix E. Marginal prior specification of differences in item difficulties

The prior distribution of the group-specific differences in item difficulty will be derived under the multiple-group IRT model. The prior for the group-specific item parameters is assumed to be multivariate normal, where the average item difficulty per group is fixed to zero due to the identification restrictions. The prior distribution of the group-specific difficulty parameters of item k given the covariance matrix $\boldsymbol{\Sigma}$ is given by

$$p\left(b_{k1},\ldots,b_{kJ} \mid \Sigma, H_1\right) \propto |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\operatorname{Tr}\boldsymbol{b}_k\boldsymbol{b}_k^t\Sigma^{-1}\right).$$

When assuming the identity matrix $\boldsymbol{I}_J$ as the hyper prior scale matrix, the Inverse Wishart prior for the covariance matrix Σ is given by

$$p(\Sigma) \propto |\Sigma|^{-\frac{2J+1}{2}} \exp\left(-\frac{1}{2}\operatorname{Tr}\Sigma^{-1}\right).$$

The marginal prior distribution of $\boldsymbol{b}_k$ is obtained by integrating out the covariance matrix Σ. It follows that

$$p(\boldsymbol{b}_k \mid H_1) \propto \int |\Sigma|^{-J-1} \exp\left(-\frac{1}{2}\operatorname{Tr}\left(\boldsymbol{b}_k\boldsymbol{b}_k^t + \boldsymbol{I}_J\right)\Sigma^{-1}\right) d\Sigma^{-1}.$$

The integral is equal to the normalizing constant of the Wishart distribution,

$$p(\boldsymbol{b}_k \mid H_1) \propto \left|\mathbf{I}_J + \boldsymbol{b}_k\boldsymbol{b}_k^t\right|^{\frac{-J-1}{2}},$$

which does not depend on the item difficulty parameters, such that the joint prior density function of the group-specific difficulty parameters is the multivariate student t with one degree of freedom and covariance matrix $\boldsymbol{I}_J$. Therefore, the joint distribution can be expressed as

$$\begin{aligned} p(\boldsymbol{b} \mid H_1) &= \frac{\Gamma((J+1)/2)}{\Gamma(1/2)\,\pi^{J/2}}\left[\mathbf{I}_J + \boldsymbol{b}_k\boldsymbol{b}_k^t\right]^{\frac{-J-1}{2}} \\ &= \frac{\Gamma((J+1)/2)}{\Gamma(1/2)\,\pi^{J/2}}\left[1 + \boldsymbol{b}_k^t\boldsymbol{b}_k\right]^{\frac{-J-1}{2}} \end{aligned}$$

where a matrix lemma (Press, 2003, p. 208) was used to obtain the kernel of the multivariate t-distribution. A linear transformation

Table 3
Average estimation accuracy results (average parameter estimates, BIAS, MSE) for five pairs of items with increasing amounts of DIF between two groups, over 50 replicated data sets.

	Cauchy Prior			Normal Prior			MML/EM		
750 subjects per group									
d_k	EAP d_k	BIAS b_{kj}	MSE b_{kj}	EAP d_k	BIAS b_{kj}	MSE b_{kj}	$\hat{d}_k$	BIAS b_{kj}	MSE b_{kj}
0.00	0.01	0.09	0.03	0.02	0.09	0.03	0.02	0.10	0.14
0.10	0.11	0.09	0.03	0.11	0.11	0.03	0.10	0.09	0.14
0.30	0.30	0.09	0.03	0.32	0.11	0.03	0.28	0.09	0.11
0.50	0.51	0.10	0.03	0.49	0.09	0.03	0.50	0.09	0.20
0.70	0.69	0.11	0.03	0.70	0.10	0.03	0.67	0.12	
									0.28
500 subjects per group									
d_k	EAP d_k	BIAS b_{kj}	MSE b_{kj}	EAP d_k	BIAS b_{kj}	MSE b_{kj}	$\hat{d}_k$	BIAS b_{kj}	MSE b_{kj}
0.00	0.00	0.12	0.04	0.00	0.11	0.05	0.00	0.12	0.07
0.10	0.00	0.10	0.04	0.00	0.12	0.04	0.07	0.11	0.08
0.30	0.15	0.12	0.05	0.18	0.13	0.05	0.29	0.13	0.08
0.50	0.56	0.12	0.04	0.57	0.14	0.06	0.48	0.14	0.10
0.70	0.92	0.11	0.04	0.91	0.12	0.04	0.67	0.13	0.10
250 subjects per group									
d_k	EAP d_k	BIAS b_{kj}	MSE b_{kj}	EAP d_k	BIAS b_{kj}	MSE b_{kj}	$\hat{d}_k$	BIAS b_{kj}	MSE b_{kj}
0.00	0.00	0.11	0.05	0.01	0.16	0.08	0.02	0.13	0.17
0.10	0.00	0.12	0.06	0.00	0.14	0.07	0.11	0.13	0.49
0.30	0.10	0.12	0.05	0.13	0.14	0.07	0.29	0.14	0.13
0.50	0.57	0.13	0.06	0.33	0.16	0.08	0.50	0.15	0.16
0.70	0.87	0.12	0.05	0.80	0.16	0.08	0.66	0.11	0.23

of the group-specific difficulty parameters is again multivariate t-distributed. Consider a contrast matrix $\boldsymbol{C}$, then $\boldsymbol{d}_k = \boldsymbol{C}\boldsymbol{b}_k$ is multivariate t-distributed with one degrees of freedom and covariance matrix $\boldsymbol{C}\boldsymbol{I}_J\boldsymbol{C}^t$.

In the situation of two groups ($J = 2$) and a linear contrast, $\boldsymbol{C} = [-1, 1]^t$, the distribution of $d_k = b_{k1} - b_{k2}$ is the univariate Student t-distribution with one degrees of freedom and scale parameter 2. Subsequently,

$$p\left(d_k \mid H_1\right) = \frac{\Gamma(1)}{\Gamma(1/2)\sqrt{2\pi}}\left[1 + d_k^2/2\right]^{-1}$$

and $p\left(d_k = 0 \mid H_1\right) = \frac{1}{\sqrt{2\pi}}$.

Appendix F. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jmp.2015.06.005.

References

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics*, *17*, 251–269.

Azevedo, C. L. N., Andrade, D. F., & 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. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459.

Bock, R. D., & Zimowski, M. F. (1997). The multiple groups IRT. In Wim J. van der Linden, & Ronald K. Hambleton (Eds.), *Handbook of modern item response theory*. Springer-Verlag.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*, 1061–1071.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A bayesian random effects model for testlets. *Psychometrika*, *64*, 153–168.

Cai, L. (2008). SEM of another flavor: Two new applications of the supplemented EM algorithm. *British Journal of Mathematical and Statistical Psychology*, *61*, 309–329.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). *IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]*. Chicago, IL: Scientific Software International.

Cowles, M. K., & Carlin, B. P. (1996). Markov chain monte carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, *91*, 883–904.

De Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*, 533–559.

De Jong, M. G., & Steenkamp, J. B. E. M. (2010). Finite mixture multilevel multidimensional ordinal IRT models for large scale cross-cultural research. *Psychometrika*, *75*, 3–32.

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

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics*, *42*, 204–223.

Flowers, L., Osterlind, S. J., Pascarella, E. T., & Pierson, C. T. (2001). How much do students learn in college?: Cross-sectional estimates using College BASE. *Journal of Higher Education*, *72*, 565–583.

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

Fox, J.-P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, *66*, 271–288.

Fox, J.-P., & Verhagen, A. J. (2010). Random item effects modeling for cross-national survey data. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-cultural analysis: Methods and applications* (pp. 467–488). London: Routeledge Academic.

Frederickx, S., Tuerlinckx, F., De Boeck, P., & Magis, D. (2010). RIM: a random item mixture model to detect differential item functioning. *Journal of Educational Measurement*, *47*, 432–457.

Glas, C. A. W., & Van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement*, *27*, 247–261.

Glas, C. A. W., van der Linden, W. J., & Geerlings, H. (2010). Estimation of the parameters in an item-cloning model for adaptive testing. In W. J. van der Linden, & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 289–314). New York: Springer.

Janssen, R., Tuerlinckx, F., Meulders, M., & De Boeck, P. (2000). A hierarchical irt model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics*, *25*, 285–306.

Jeffreys, H. (1961). *Theory of probability, 3rd*. Oxford: Oxford University Press.

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

Kim, S.-H., Cohen, A. S., & Park, T.-H. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement*, *32*, 261–276.

Langer, M. (2008). A reexamination of Lord's Wald test for differential item functioning using item response theory and modern error estimation (unpublished doctoral dissertation). University of North Carolina, Chapel Hill.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Muthén, B., & Asparouhov, T. New Features in Mplus v7 Lecture 3 (2013a). Available online at: https://www.statmodel.com/examples/webnotes/webnote17.pdf [Accessed 26.11.13].

Muthén, B., & Asparouhov, T. *New Methods for the Study of Measurement Invariance with Many Groups*. (2013b) Available online at: http://www.statmodel.com/download/PolAn.pdf [Accessed 26.11.13].

Muthén, L. K., & Muthén, B. O. (2012). *Mplus (Version 7)*. Los Angeles, CA: Muthén and Muthén.

Osterlind, S. J., Robinson, R. D., & Nickens, N. M. (1997). Relationship between collegians' perceived knowledge and congeneric tested achievement in general education. *Journal of College Student Development*, *38*, 255–265.

Patz, R. J., & Junker, B. W. (1999a). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, *24*, 342–366.

Patz, R. J., & Junker, B. W. (1999b). A straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*, 146–178.

Press, S. J. (2003). *Subjective and objective Bayesian statistics*. New York: Wiley.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & review*, *16*, 225–237.

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *The American Statistician*, *55*, 62–71.

Sinharay, S., Johnson, M. S., & Williamson, D. M. (2003). Calibrating item families and summarizing the results using family expected response functions. *Journal of Educational and Behavioral Statistics*, *28*, 295–313.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale, NJ: Lawrence Erlbaum.

Verdinelli, I., & Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage–Dickey density ratio. *Journal of the American Statistical Association*, *90*.

Verhagen, A. J., & Fox, J.-P. (2013a). Bayesian tests of measurement invariance. *The British Journal of Mathematical and Statistical Psychology*,.

Verhagen, A. J., & Fox, J.-P. (2013b). Longitudinal measurement in health-related surveys. A Bayesian joint growth model for multivariate ordinal responses. *Statistics in Medicine*,.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. P. P. P. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive psychology*, *60*, 158–189.

Wagenmakers, E.-J., Verhagen, J., Ly, A., Bakker, M., Lee, M. D., Matzke, D., Rouder, J. N., & Morey, R. D. (2014). A power fallacy. *Behavior Research Methods*, 1–5.

Wagenmakers, E.-J., Verhagen, J., Ly, A., Matzke, D., Steingroever, H., Rouder, J.N., & Morey, R.D. The need for Bayesian hypothesis testing in psychological science. In S. O. Lilienfeld & I. Waldman (Eds.), Psychological science under scrutiny: Recent challenges and proposed solutions. John Wiley and Sons, (in press).

Wetzels, R., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2010). An encompassing prior generalization of the Savage–Dickey density ratio test. *Computational Statistics & Data Analysis*, *54*, 2094–2102.

Wetzels, R., Raaijmakers, J., Jakab, E., & Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian *t*-test. *Psychonomic Bulletin & Review*, *16*, 752–760.

Woods, C. M., Cai, L., & Wang, M. (2012). The Langer-improved Wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT. *Educational and Psychological Measurement*, *73*, 532–547.