Testing and Explaining Differences in Common and Residual Factors Across Many Countries

Jak, S.

DOI
10.1177/0022022116674599

Publication date
2017

Document Version
Final published version

Published in
Journal of Cross-Cultural Psychology

Journal of Cross-Cultural Psychology
2017, Vol. 48(1) 75–92

DOI: 10.1177/0022022116674599
jccp.sagepub.com
SAGE

Testing and Explaining Differences in Common and Residual Factors Across Many Countries

Suzanne Jak[1]

Abstract

To make valid comparisons across countries, a measurement instrument needs to be measurement invariant across countries. The present article provides a nontechnical exposition of a recently proposed multilevel factor analysis approach to test measurement invariance across countries. It is explained that strong factorial invariance across countries implies equal factor loadings across levels and zero residual variance at the country level in a two-level factor model. Using two-level factor analysis, the decomposition of the variance at each level can be investigated, measurement invariance can be tested, and country-level variables can be added to explain differences in the common or residual factors. The approach is illustrated using two examples. The first example features data about well-being from the European Social Survey and the second example uses data about mathematical ability from the Programme for International Student Assessment (PISA) study. The input-files and annotated output-files for both examples are provided in the supplementary files.

Keywords

measurement invariance, factorial invariance, multilevel factor analysis, cross-cultural invariance

When comparing psychological test scores across countries, it is important that these measurements are invariant across countries (see Billiet, 2003; Byrne & van de Vijver, 2010; Davidov, Meuleman, Cieciuch, Schmidt, & Billiet, 2014; Poortinga, 1989). Suppose researchers want to investigate across-country differences in well-being and use six items to operationalize well-being. If the measurements function identically in all countries, all persons who have the same true value on well-being will have the same expected items scores, regardless of the country they live in. If this holds, then the measurements of well-being are invariant across countries (Mellenbergh, 1989). If the measurements are not invariant, other variables than well-being have systematic influence on the item scores, rendering comparisons across countries biased.

Researchers are increasingly aware of the advantages of structural equation modeling (particularly factor analysis) over the use of composite scores, as the common factors are free of measurement error (Bollen, 1989). To clarify the different parameters involved in a factor model, Figure 1 shows a graphical representation of a one-factor model with four indicators.

[1]University of Amsterdam, The Netherlands

Corresponding Author:
Suzanne Jak, Methods and Statistics, Research Institute of Child Development and Education, University of Amsterdam, Nieuwe Achtergracht 127, 1018 WS, Amsterdam, The Netherlands.
Email: S.Jak@uva.nl

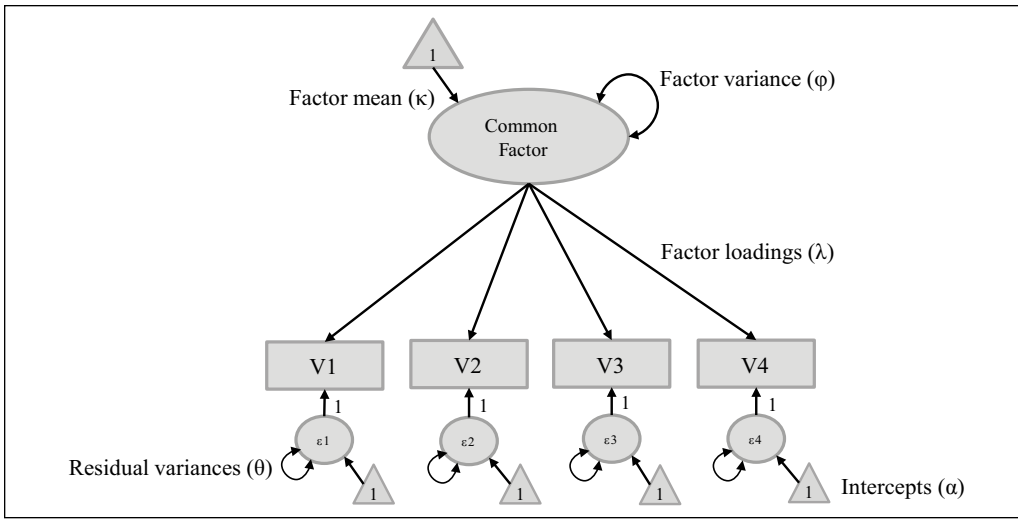

Figure 1. A one-factor model on four items.
Note. Observed variables are represented by squares. Latent variables are represented by ellipses. Single headed arrows indicate regression coefficients. Double headed arrows represent variances (or covariances). The effects of residual factors on indicators are fixed at 1 by default. The small triangles represent constants of 1. The regression of the common factor on this constant represents the factor mean. The regressions of the residual factors (ε_1-ε_4) on the constant of 1 represent the residual means or intercepts.

Measurement invariance is most often tested using multigroup confirmatory factor analysis (Sörbom, 1974). Within multigroup factor analysis, a distinction is made between four levels of invariance—being configural invariance, weak factorial invariance, strong factorial invariance, and strict factorial invariance. Adequate comparisons of factor means across countries are possible if strong factorial invariance across countries holds (Meredith, 1993; Widaman & Reise, 1997). Strong factorial invariance across countries comprises equality of factor loadings and intercepts across countries. If the intercepts differ across countries, but the factor loadings do not, then weak factorial invariance holds, allowing comparisons of variances and regression coefficients across countries, but not of factor means.[1] If the pattern of factor loadings is the same across countries, but the values are different, then configural invariance holds. Configural invariance does not allow for meaningful comparisons of factors in multigroup modeling. Strict factorial invariance is needed if observed scores (as opposed to factor means) will be compared across groups. Strict factorial invariance comprises equal residual variances in addition to equal factor loadings and intercepts across groups. However, most researchers focus on comparisons of factor means, for which strong factorial invariance is sufficient.

When researchers are interested in differences between relatively large numbers of countries, multilevel structural equation modeling (MLSEM) is a useful statistical technique. Recently, Jak, Oort, and Dolan (2013) showed how strong factorial invariance across groups in a multigroup factor model implies equal factor loadings across levels and zero residual variance at the between level in a two-level factor model. The present article serves as a nontechnical introduction to testing strong factorial invariance across countries using multilevel factor models. Jak et al. called the proposed framework a test on cluster bias. However, when applied to countries as clusters, a test on country bias may be a more appropriate term.

MLSEM has been promoted in cross-cultural psychology before (Cheung, Leung, & Au, 2006; Davidov, Dülmer, Schlüter, Schmidt, & Meuleman, 2012; Selig, Card, & Little, 2008), as it provides an excellent framework for the investigation of within- and between-country differences at the appropriate levels of analysis. Moreover, the use of multilevel factor analysis

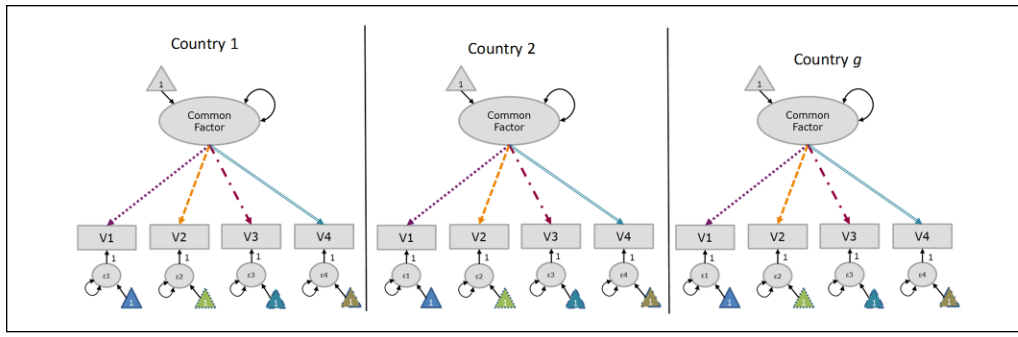

Figure 2. Strong factorial invariance across countries in a multigroup model.
Note. Factor loadings and intercepts with the same dash-type of line are equal across countries. When fitting this model, for identification, the factor mean and variance needs to be fixed at, respectively, 0 and 1 in one group, and can be freely estimated in the other groups (Millsap & Yoon, 2007).

facilitates the investigation of the relation of latent variables (as opposed to composite scores) with observed country-level variables like general income level or literacy level.

Davidov et al. (2012) presented a procedure to test factorial invariance across countries using multigroup factor analysis, after which they use MLSEM to explain the bias across countries. The researchers stress the importance of explaining possible bias using country-level variables, as such an analysis will give deeper insight into the differences across countries. They showed, for example, that a country's level of human development explained measurement bias in an item measuring "universalism." The procedure of Jak et al. that I illustrate in this article partly overlaps with the procedure of Davidov et al. The main difference is that I use MLSEM throughout the procedure, and use a top-down approach (starting with the most restrictive model), while Davidov et al. use multigroup factor analysis to explicitly test each step of factorial invariance in a bottom-up approach (starting with the least restrictive model).

In the next section, the test on country bias as proposed by Jak et al. will be explained. In addition, it will be shown how factor variance may be decomposed into a within- and a between-countries part, and how additional country-level variables may be included in the two-level factor model. Next, the procedure is illustrated with two example data sets from the *European Social Survey* (ESS, 2014) and PISA (Organisation for Economic Co-Operation and Development [OECD], 2013).

Testing Strong Factorial Invariance Across Countries

Figure 2 gives a graphical overview of a one-factor model with four indicators, representing strong factorial invariance across *g* countries. Assuming that the factor structure is equal across countries (i.e., that *configural* invariance holds), one could use multigroup factor analysis to test the equality of factor loadings and intercepts across groups. However, this may be a cumbersome strategy when the number of countries involved is large (Muthén & Asparouhov, 2013). Jak et al. (2013) showed that when strong factorial invariance across clusters (countries) holds, this condition translates to a two-level factor model with equal factor loadings across levels, and no residual variance at the between (country) level. In cross-cultural research, the equality of concepts across levels is called *isomorphism* (van de Vijver, van Hemert, & Poortinga, 2008). Specifically, equality of factor loadings across levels in MLSEM is denoted *strong metric isomorphism* by Tay, Woo, and Vermunt (2014). If strong metric isomorphism holds, then the common factors have the same interpretation across levels. In the example of well-being, this means that the factor at the country level represents countries' average well-being, and the factor at the within level represents individual deviations from

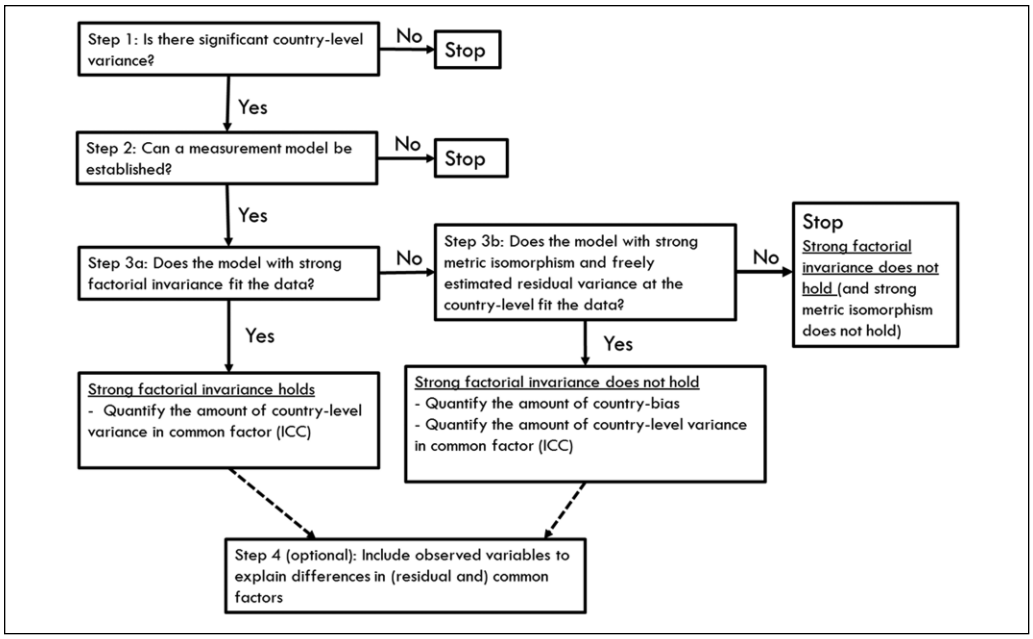

Figure 3. Flowchart of the suggested procedure.
Note. ICC = intraclass correlation.

the respective countries' average well-being. The factor well-being is thus decomposed into a within-country and a between-country part. Consequently, one can evaluate how much of the common factor variance exists between countries and how much exists within countries, and one could compare structural relations between variables across levels. If strong metric isomorphism does not hold, then the common factor at the between level has a different interpretation than the common factor at the within level, and one cannot validly compare results across levels (Guenole, 2016). Equality of factor loadings across levels is a prerequisite for strong factorial invariance across countries. All factor models in the proposed framework therefore assume that factor loadings are equal across levels, that is, that strong metric isomorphism holds. I refer to Tay et al. for a discussion of weaker forms of isomorphism.

In addition to strong metric isomorphism, strong factorial invariance across countries implies zero residual variance at the country level.[2] This means that all observed differences between countries are differences in the common factors. If there are other variables than the common factor causing country differences in the indicators, their (biasing) influence will show up as residual variance at the country level. Van de Vijver et al. (2008) described a taxonomy of multilevel models with varying restrictions in cross-cultural data. They state,

> Conceptually, there is a close relationship between equivalence across cultures and across levels. Scores of individuals in two cultures cannot be taken to be equivalent when an additional variable has to be introduced to account for differences at the culture level. (Van de Vijver et al., 2008, p. 21)

Translated to the two-level model, it means that if there are country-level differences that the common factors do not account for, there is measurement bias across countries (or equivalently, strong factorial invariance across countries does not hold). One could follow the following four steps to test strong factorial invariance across countries, and to quantify and explain possible differences between countries. A flowchart of the suggested procedure is provided in Figure 3.

Step 1: Testing the Existence of Variance at the Country Level

As a first step before fitting two-level factor models, one may test whether two-level modeling is actually appropriate. Obviously, if there are no differences between countries on the observed variables, fitting a model to explain the country-level differences does not make sense. A useful check on the necessity of multilevel modeling is to fit a so-called *saturated model* at the within level (a model where all variables are correlated with each other), and to restrict all variances and covariances to be zero at the between level (a so-called *null model*). The saturated model always fits the data perfectly, so any significant misfit of this model indicates that there is significant variance at the between level (Hox, 2002; Muthén, 1994). The amount of between-level variance can be quantified by the intraclass correlations (ICC) of the observed variables. The ICC is interpreted as the proportion of the total variance that exists at the between level.

Step 2: Finding an Appropriate Factor Structure Across Countries

Before testing factorial invariance, one should establish the common factor structure across countries. There are several ways to do this. One method is to decide on the factor structure by modeling the within factor structure only (Jak, Oort, & Dolan, 2014). The researcher then fits the proposed factor model at the within level, and a saturated model at the between level.[3] If no adequate factor structure can be found, one should not continue with the next steps of the analysis.

Step 3a: Fitting a Two-Level Factor Model Representing Strong Factorial Invariance

The next step is to fit the model with equal factor loadings across the within and between-country level and zero residual variance at the country level to the data. If this model fits the data, one can conclude that strong factorial invariance across countries holds. The equal factor loadings across levels imply that the common factor at the country level captures the country-mean differences of the factor within the countries (Mehta & Neale, 2005). One can calculate the proportion of factor variance that is present at the country level using the factor variances (φ), and the formula for the ICC: $\varphi_{between} / (\varphi_{between} + \varphi_{within})$, as explained by Mehta and Neale (2005). The variance decomposition of the common factor(s) may be interesting in itself, as it indicates how much of the construct is culturally determined (see, for example, Fischer & Schwartz, 2011).

If strong factorial invariance is not rejected, and one has observed country-level variables that can be used to explain differences in the common factor, one may continue to Step 4. If the model at Step 3a does not fit the data, strong factorial invariance across countries is rejected. One may then continue to Step 3b.

Step 3b: Accounting for Bias by Estimating Residual Variance at the Country Level

If strong factorial invariance is rejected at Step 3a, the next step is to account for country bias by freely estimating the residual variance at the country level. If this model fits the data, we conclude that there is country bias present in one or more of the indicators. In this model, the differences in the common factors across countries are represented by the common factor at the country level, and all remaining structural differences in the item scores between the countries are represented by the residual variance (Muthén, 1990; Rabe-Hesketh, Skrondal, & Pickles, 2004).

The significance of the residual variance can be evaluated using Wald tests (based on the standard errors) or using likelihood ratio tests (comparing the model fit of a model with and

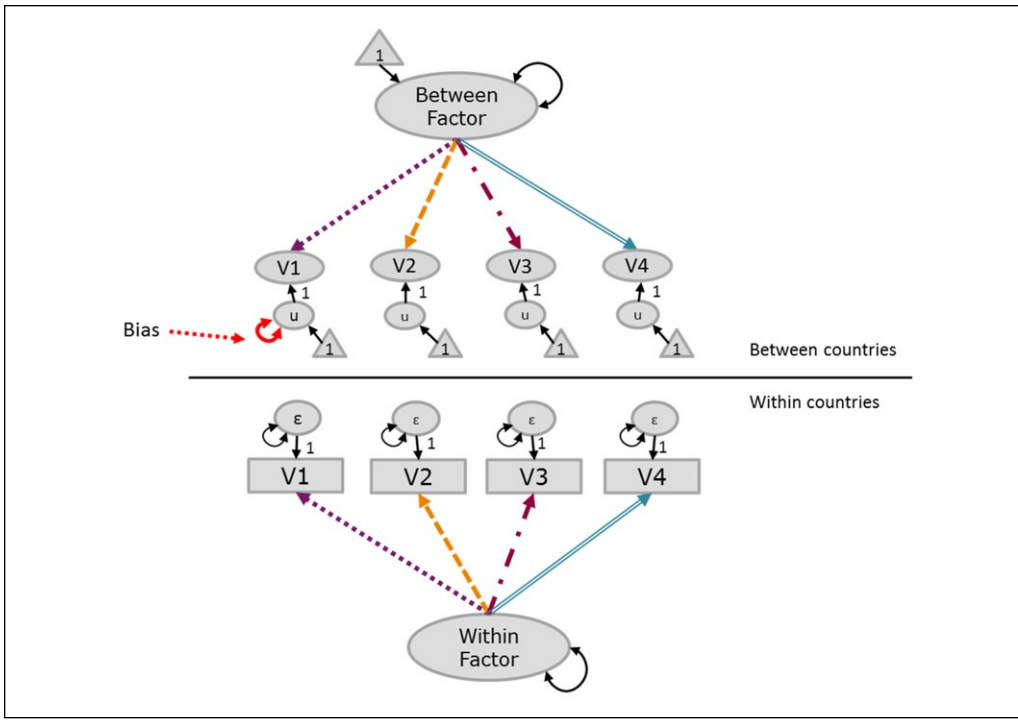

Figure 4. Strong factorial invariance across countries in a two-level model (with country bias in Item 1).
Note. Factor loadings with the same dash-type of line are equal across levels. In a multilevel model, the within part is all mean-centered, so all means are zero. The between part of the indicators is represented by small ellipses. For identification, the factor variance at one of the levels needs to be fixed at 1, with equal factor loadings, the variance at the other level can be freely estimated.

without estimated residual variance at the between level). If residual variance for one or more indicators is found to be nonzero, then strong factorial invariance is rejected, but partial strong factorial invariance may still hold (Byrne, Shavelson, & Muthén, 1989). That is, country bias may be present in only a part of the items. Although not often observed with real data, it is possible that country bias in several items is caused by the same biasing factor, which would lead to significant residual *covariance* at the between level (Jak et al., 2013).

Figure 4 shows a graphical representation of the one-factor model on four indicators, with country bias in Item 1. Simulation research by Jak et al. (2013) showed that both differences in factor loadings across countries and differences in intercepts across countries show up as residual variance in the two-level model with strong metric isomorphism. This means that detecting significant residual variance implies that strong factorial invariance across countries does not hold, but it does not give information about the source of bias (differences in intercepts and/or differences in factor loadings).

If there is significant residual variance at the between level, representing country bias, it may be of interest to calculate how large the proportion of bias is (see Jak, 2014). Two proportions may be of interest: the proportion of measurement bias in the country-level variance and the proportion of measurement bias in the total variance (within + between country variance). In a factor model, the total variance at a certain level for one item can be calculated using the level-specific parameter estimates related to the item. For example, the total variance within countries in one item, loading on one factor is,

$$\lambda_W^2 \varphi_w + \theta_w, \tag{1}$$

where λ represents a factor loading, φ represents common factor variance and θ represents residual variance, and subscript "w" indicates "within-level." At the between-country level (indicated by subscript "b"), this would be,

$$\lambda_b^2 \phi_b + \theta_b. \tag{2}$$

Therefore, the proportion of measurement bias (residual variance) in the total variance is calculated as,

$$\frac{\theta_b}{\left(\lambda_b^2 \varphi_b + \theta_b + \lambda_w^2 \varphi_w + \theta_w\right)}. \tag{3}$$

If the model at Step 3b does not fit the data, this can be caused by factor loadings being unequal across levels (leading to different interpretations of the factors across levels), or even due to unequal factor structures across levels. In any case, an ill-fitting model at this step indicates that both strong factorial invariance across countries and strong metric isomorphism as defined by Tay et al. (2014) do not hold. This situation is not further discussed here, but see Tay et al. for a discussion of less strict levels of isomorphism.

If the model with equal factor loadings and freely estimated residual variance fitted the data, and one has observed country-level variables that can be used to explain bias or to explain differences in the common factor, one may continue to Step 4.

Step 4: Explaining Country-Level Differences in Common and Residual Factors

The common factor at the between level represents country-level differences in the common factor means. Other country-level variables can be added to the model to explain these differences, as suggested by Davidov et al. (2012). For example, country differences in well-being may be explained by the economic situation of countries (see Dolan, Peasgood, & White, 2008), country differences in mathematical ability may be explained by cultural differences in number vocabulary (see Göbel, Shaki, & Fischer, 2011), and many other examples can be thought of.

If measurement bias was found in Step 3b, researchers should try to interpret the bias (Ackerman, 1992). The question to consider in this case is, "What other variables than the common factor could lead to country differences in this item?" If one has a measure of the potential explanation of the bias, one can include them in the between-level model to explain the bias away (Davidov et al., 2012). There is, however, a restriction in which part of the residual variance we can explain using standard MLSEM. Regressing residual factors at the between level on observed variables directly lead to possible explanations of *uniform bias* (differences in intercepts), whereas *nonuniform bias* (differences in factor loadings) may be explained by regressing the residual factors at the between level on the interaction of the common factor and observed variables at the between level. Recent developments using latent interaction terms or moderated factor analysis provide a viable method to explain nonuniform bias (Barendse, Oort, & Garst, 2010; see also Molenaar, Dolan, Wicherts, & van der Maas, 2010), but have not been tested at the between level in the multilevel situation. Therefore, the current approach only considers explaining uniform bias across countries.

Illustrations

The four steps of the suggested procedure will be illustrated using two examples of cross-national data. As my intention is only to illustrate the approach, and not to test specific substantial

hypotheses, I do not focus on the theoretical background of the variables under consideration. However, I will state which research questions can be answered by applying the proposed techniques to the specific data sets. The first example shows how measurement invariance and country-level differences can be investigated on six items measuring well-being. The second example illustrates the approach on nine items measuring mathematical ability.

Example 1—Emotional Well-Being

I will illustrate the testing procedure using six items to measure "emotional wellbeing" that were included in round 2012 of the ESS (2014; Huppert et al., 2009).[4] These items originally stem from a depression scale for research in the general population (Center for Epidemiologic Studies Depression Scale [CES-D], Radloff, 1977). Three items are positively formulated, asking how often in the last week a respondent was happy (WRHPP), enjoyed life (ENJLF), and felt calm and peaceful (FLTPCFL). The other three items were negatively phrased, asking how often in the last week a respondent felt depressed (FLTDP), felt sad (FLTSD), and felt anxious (FLTANX). The items were scored on a 4-point Likert-type scale ranging from *none or almost none of the time* to *all or almost all of the time*. I treat the responses to the 4-point scale as approximately continuous (Example 2 shows how to handle discrete responses). Round 2012 of the ESS included data from 54,673 respondents from 29 countries on these items. The amount of missing responses ranged from 0.69% to 1.28% per variable. In addition to the well-being items, I downloaded two variables at the country level, being the variable Individual Liberties (IL) from the Democracy barometer from 2012 and the Corruption Perceptions Index (CPI) score from 2013. Higher scores on IL reflect more perceived individual liberties in a country. Higher scores on CPI reflect less perceived corruption in a country. Further details about these data and the data itself can be freely downloaded from http://www.europeansocialsurvey.org/, using the variable labels provided above. I provide the syntax to fit the models and annotated output in the supplementary material.

Research questions. The research questions that can be answered using these data are as follows:

> **Research Question 1:** Does strong factorial invariance across countries of these emotional well-being items hold?
> **Research Question 2:** How much of the total variance in people's well-being is at the country level?
> **Research Question 3:** Can the contextual variables IL and CPI explain (part of) the bias and/or common factor variance at the country level?

Analysis. All models were fit to the data with Mplus Version 7 (Muthén & Muthén, 1998-2015), using maximum likelihood estimation (MLR). This estimation method provides a test statistic that is asymptotically equivalent to the Yuan–Bentler T2 test statistic (Yuan & Bentler, 2000), and standard errors that are robust for nonnormality.[5] Statistical significance of the χ^2 statistic indicates that exact fit of the model has to be rejected. With large sample sizes, very small model misspecifications may lead to rejection of the model. Therefore, in addition to the adjusted χ^2 statistic, I consider two measures of approximate fit—the root mean square error of approximation (RMSEA; Steiger & Lind, 1980) and the comparative fit index (CFI; Bentler, 1990). RMSEA values smaller than .05 indicate close fit and values smaller than .08 are considered satisfactory (Browne & Cudeck, 1992). CFI values more than .95 indicate reasonably good fit (Hu & Bentler, 1999). Statistical significance of parameter estimates was evaluated using univariate Wald tests given the standard errors. To decrease the risk of capitalizing on chance when testing several hypotheses simultaneously, I use alpha levels of .01 instead of .05 when testing for the

significance of residual variances at the country level.[6] Maximum likelihood estimation fits the model using all available information at the individual level, so no deletion or imputation of missing data was needed.

Results. The results will be presented following the four steps described before.

Step 1: Testing the existence of variance at the country level. Fitting a saturated model (all variables correlated with each other) at the within level and a null model (all variance fixed at 0) at the between level lead to a significant chi-square statistic: $\chi^2(21) = 11{,}848.43$, $p < .05$. This indicates that there is significant variance at the country level. The intra class correlation of the items ranged from .035 to .122.

Step 2: Finding an appropriate factor structure across countries. As the six well-being items are treated as measuring one dimension by the ESS, I first fitted a one-factor model at the within level, with a saturated model at the between level. This model did not fit the data well: $\chi^2(9) = 2{,}999.82$, $p < .05$, RMSEA = .078, CFI = .86. A sensible alternative model is a two-factor model, where the negative items load on a negative well-being factor, and the positive items load on a positive well-being factor. This model fitted the data well: $\chi^2(8) = 587.562$, $p < .05$, RMSEA = .036, CFI = .97. Although the positive and negative factor are quite strongly correlated ($r = -.69$), the dimension analysis shows that they are two separate dimensions. Therefore, the two-factor model was used as the measurement model in subsequent steps.

Step 3a: Fitting a two-level factor model representing strong factorial invariance. A two-level model with equal factor loadings and zero residual variance at the country level did not fit the data, $\chi^2(26) = 8{,}196.16$, $p < .05$, RMSEA = .076, CFI = .62. Hence, strong factorial invariance across countries was rejected.

Step 3b: Accounting for bias by estimating the residual variance at the country level. The two-level model with equal factor loadings across levels and freely estimated residual variance at the between level shows good fit to the data: $\chi^2(20) = 1{,}145.3$, $p < .05$, RMSEA = .031, CFI = .95. The parameter estimates and standard errors of the fitted model can be found in Table 1. I evaluated the significance of the between-level residual variance using the provided standard errors. Significant residual variance was present in the items "enjoy life," "feel calm and peaceful," and "felt anxious," but not in the other three items, indicating partial invariance (Byrne et al., 1989) for both factors. Apparently, there are other factors influencing the country-level scores on these three items than the common factors of well-being. The proportions of country-level bias at the between level in these items were, respectively, .361, .468, and .591. Overall, the bias made up a proportion of, respectively, .019, .020, and .068 of the total variance. There was no significant residual covariance between the three items at the between level.

We can calculate the proportion of variance in negative and positive well-being at the country level using the formula for the ICC: $\varphi_{between} / (\varphi_{between} + \varphi_{within})$. In this example, the ICC of the positive well-being factor is .062 / (.062 + 1) = .058. So, 5.8% of the variance in positive well-being is at the between-country level, and 94.2% is at the within-country level. The ICC of the negative well-being factor is .137 / (.137 + 1) = .120. So, 12.0% of the variance in negative well-being is at the between-country level, and 88.0% at the within-country level.

Step 4: Explaining differences in common and residual factors across countries by a contextual variable. The individual liberties score (IL) and the CPI were added to the country-level factor model. The fit of this model was good: $\chi^2(28) = 1{,}383.1$, $p < .05$, RMSEA = .030, CFI = .95. I tested whether IL and CPI could explain part of the bias in the items "enjoy life," "feel calm and peaceful," and "felt anxious" by estimating direct effects of the contextual variables on the items.

Table 1. Parameter Estimates (Est.), Standard Errors (*SE*), *p* Values (*p*) and Standardized Parameter Estimates (Std.) for the Two-Factor Model on Six Well-being Items With Across-Level Invariance.

	Within level				Between level					
Parameter	Est.	*SE*	*p*	Std.	Est.	*SE*	*p*	Std.	Bias[a]	Bias[b]
Factor loadings										
Positive well-being										
λ_{11}	0.626	0.013	<.01	0.777	0.626	0.013	<.01	0.894		
λ_{21}	0.632	0.011	<.01	0.756	0.632	0.011	<.01	0.801		
λ_{31}	0.507	0.012	<.01	0.621	0.507	0.012	<.01	0.727		
Negative well-being										
λ_{42}	0.510	0.013	<.01	0.740	0.510	0.013	<.01	0.949		
λ_{52}	0.517	0.015	<.01	0.754	0.517	0.015	<.01	0.949		
λ_{62}	0.444	0.017	<.01	0.627	0.444	0.017	<.01	0.640		
Factor variance										
$\varphi_{positive}$	1.000	—	—	1.000	0.062	0.015	<.01	1.000		
$\varphi_{negative}$	1.000	—	—	1.000	0.137	0.034	<.01	1.000		
Factor covariance										
φ_{21}	−0.694	0.012	<.01	−0.694	−0.080	−0.020	<.01	−0.871		
Residual variance										
θ_{11}	0.258	0.010	<.01	0.397	0.006	0.002	.014	0.201		
θ_{22}	0.299	0.015	<.01	0.428	0.014	0.005	<.01	0.358	.361	.019
θ_{33}	0.411	0.014	<.01	0.615	0.014	0.004	<.01	0.472	.468	.020
θ_{44}	0.216	0.011	<.01	0.453	0.004	0.003	.181	0.100		
θ_{55}	0.203	0.007	<.01	0.432	0.004	0.004	.272	0.100		
θ_{66}	0.304	0.016	<.01	0.607	0.039	0.010	<.01	0.591	.591	.068

Note. Items 1 to 6 are (a) happy, (b) enjoy, (c) calm, (d) depressed, (e) sad, (f) anxious.
[a]Proportion of country-level bias in country-level variance: For Item 2, .014 / ($.632^2$ × .062 + .014) = .361; for Item 3, .014 / ($.507^2$ × .062 + .014) = .468, and for Item 6, .039 / ($.444^2$ × .137 + .039) = .591.
[b]Proportion of country-level bias in total variance, calculated by adding the within-level variance to the denominator: For Item 2, .014 / ($.632^2$ × .062 + .014+ $.632^2$ + .299) = .019; for Item 3, .014 / ($.507^2$ × .062 + .014 + $.507^2$ + .411) = .020; and for Item 6, .039 / ($.444^2$ × .137 + .039 + $.444^2$ + .304) = .068.

One of these direct effects was statistically significant, which was the effect of CPI on the item "enjoy life" (β = .31). CPI thus explains part of the measurement bias across countries in this item. Specifically, for equal levels of well-being, countries that have higher scores on CPI report higher scores on "enjoy life."

A graphical representation of the model with this direct effect and standardized parameter estimates is given in Figure 5. All correlations between the two contextual variables and the factors were significantly different from zero and in the expected directions. IL and CPI were positively correlated (r = .39). Positive well-being was positively correlated with IL (r = .47) and with CPI (r = .69). Negative well-being was negatively correlated with IL (r = −.45) and with CPI (r = −.91).

Conclusion and Discussion. The well-being items from the ESS in 2012 are partially invariant across countries. Two of the positive well-being items were biased, indicating that given equal levels of the common factor positive well-being, there are still country-level differences in the items "enjoy life" and "feel calm and peaceful," whereas this is not the case for "feel happy." Given that "feeling happy" seems to be a more general statement than "feel calm and peaceful," it may make sense that countries differ more with regard to the more specific statements. Statistically, part of the bias in "enjoy life" could be explained by country level's perceived corruption.

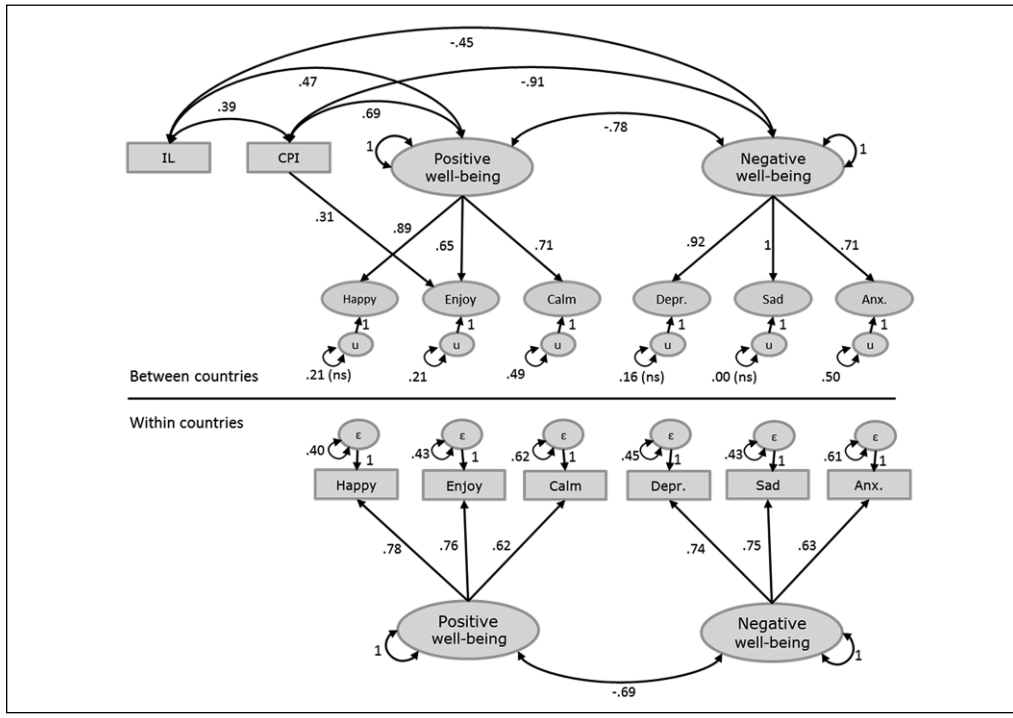

Figure 5. Final two-level factor model on well-being items with the contextual variables Individual Liberties (IL) and the Corruption Perception Index (CPI), with standardized parameter estimates.
Note. (ns) indicates that the parameter estimate was not significantly different from zero. All parameter estimates are completely standardized within the respective level. In the unstandardized solution, the factor loadings are equal across levels. Means are not represented in the model.

The negative item with bias was "anxious," so given equal levels of negative well-being, people from different countries differ in their levels of anxiety, but not in sadness or depression. A possible explanation could be that anxiety may be more influenced by external factors than sadness and depression. So although IL and CPI did not explain the bias, there may be other variables influencing people's anxiety.

Example 2—Mathematical Ability

The second data on which I will illustrate the presented approach are data collected by the PISA study in 2012. PISA collects data from 15-year-old student's abilities in reading, mathematics, and science across a large number of countries (OECD, 2013). In PISA, not all students answer all items, but the items are subdivided into "booklets," which are randomly assigned to students. I selected students who received the first booklet, and from this booklet, I selected only those items that were scored as "no credit" or "full credit" (as opposed to "partial credit"). When two or more items were subquestions within a larger question (called "units" by PISA), I selected only the first item. This way, I ended up with the responses on nine items measuring mathematical ability, from 17,493 students in 36 countries. All items are worded math problems. For example, in Item 8, a picture and description of a revolving door is presented, and students have to calculate the size in degrees of the angle between two door wings. In Item 6, the students are provided a recipe for sauce, and then have to calculate the needed amount of one ingredient for a certain quantity of sauce. Not all questions are freely available to the public, as they may be used

in future PISA rounds. Therefore, we cannot know the exact content of Items 2, 4, 7, and 9. The amount of missing responses ranged from 0.64% to 2.25% per variable. As contextual variables, I aggregated the variable "shortage of math teachers in the school" and the variable "ratio of student to math teachers" to the country level. It could be expected that mathematical ability is negatively related with both shortage of teachers and with the student to teacher ratio. The syntax to fit the models, as well as annotated output, is provided in the supplementary material.

Research questions. The research questions that can be answered using these data are as follows:

Research Question 1: Is the measurement of mathematical ability with these items measurement invariant across countries?
Research Question 2: How much of the total variance in mathematical ability is at the country level?
Research Question 3: Can the contextual variables shortage of math teachers and student–teacher ratio explain (part of) the bias and common factor variance at the country level?

Analysis. The item responses were dichotomously scored (incorrect/correct), while factor analysis is originally developed for continuous variables. To model the discrete observations, I estimated all models using the multilevel version of diagonally weighted least squares (WLSMV in Mplus; Asparouhov & Muthén, 2007). This approach assumes that underlying the observed dichotomous response, there is an unobserved latent continuous variable. In this case, underlying each dichotomous item, there is a continuous variable representing item-specific mathematical ability. Once some (to be estimated) threshold value on the underlying unobserved variable is crossed, one will give the correct response, and otherwise one will give the incorrect response. The threshold values are parameters in the model, and are essentially similar to the (inversed) difficulty parameter in item response theory models. The estimated factor loadings can be viewed as the discrimination parameters in item response models (Takane & de Leeuw, 1987). An alternative estimator to WLSMV would be MLR in Mplus. However, MLR with dichotomous data is very computational intensive, and often leads to problems (Grilli & Rampichini, 2007). WLSMV-estimation essentially replaces complex model estimation with high-dimensional numerical integration by the estimation of multiple smaller models with low-dimensional numerical integration. With WLSMV, missing data are handled using pairwise deletion, so each correlation between the underlying response variables is estimated using the information that is available for the two variables under consideration.

Similar to the previous example, we evaluate the chi-square, RMSEA, and CFI to judge the fit of the models, and I will use an alpha level of .01 when testing for country bias.

Results

Step 1: Testing the existence of variance at the country level. The two-level model with a saturated model at the within level and a null model at the between level lead to a significant chi-square statistic, $\chi^2(45) = 200.78$, $p < .05$, indicating that significant variance exists at the country level. The ICCs of the items ranged from .057 to .138.

Step 2: Finding an appropriate factor structure across countries. All items were worded math problems, and they were expected to measure one dimension. I fitted a one-factor model at the within level, with a saturated model at the between level. This model fitted the data well: $\chi^2(27) = 36.74$, $p = .10$, RMSEA = .005, CFI = .998. The one-factor model was retained for the further analyses.

Step 3a: Fitting a two-level factor model representing strong factorial invariance. The model representing strong factorial invariance across countries showed good fit to the data, $\chi^2(71) = 84.64$,

Table 2. Parameter Estimates (Est.), Standard Errors (*SE*), *p* Values (*p*), and Standardized Parameter Estimates (Std.) for the One-Factor Model on Nine Mathematical Ability Items With Strong Factorial Invariance.

	Within level					Between level			
Parameter	Est.	*SE*	*p*	Std.	Parameter	Est.	*SE*	*p*	Std.
Factor loadings									
λ_{11}	0.921	0.025	<.01	0.680		0.921	0.025	<.01	1.064
λ_{21}	0.756	0.026	<.01	0.602		0.756	0.026	<.01	0.785
λ_{31}	1.053	0.033	<.01	0.724		1.053	0.033	<.01	0.772
λ_{41}	0.561	0.027	<.01	0.485		0.561	0.027	<.01	0.729
λ_{51}	1.012	0.062	<.01	0.711		1.012	0.062	<.01	0.889
λ_{61}	0.986	0.026	<.01	0.702		0.986	0.026	<.01	1.004
λ_{71}	0.413	0.027	<.01	0.378		0.413	0.027	<.01	0.655
λ_{81}	1.113	0.041	<.01	0.744		1.113	0.041	<.01	0.946
λ_{91}	0.687	0.029	<.01	0.565		0.687	0.029	<.01	0.697
Factor variance									
φ_{math}	1.000	—	—	1.000		0.182	0.038	<.01	1.000
Residual variance					Thresholds				
θ_{11}	1.000	—	—	0.538	τ_1	0.211	0.066	<.01	—
θ_{22}	1.000	—	—	0.637	τ_2	1.753	0.066	<.01	—
θ_{33}	1.000	—	—	0.476	τ_3	0.838	0.117	<.01	—
θ_{44}	1.000	—	—	0.764	τ_4	−1.447	0.055	<.01	—
θ_{55}	1.000	—	—	0.495	τ_5	1.483	0.047	<.01	—
θ_{66}	1.000	—	—	0.508	τ_6	−0.437	0.072	<.01	—
θ_{77}	1.000	—	—	0.857	τ_7	−0.828	0.085	<.01	—
θ_{88}	1.000	—	—	0.446	τ_8	−0.343	0.085	<.01	—
θ_{99}	1.000	—	—	0.681	τ_9	−0.770	0.085	<.01	—

$p = .12$, RMSEA = .003, CFI = .997. Hence, it is concluded that strong factorial invariance across countries holds for these variables. Table 2 shows the parameter estimates and standard errors of the fitted model.

Calculating the proportion of the variance in the common factor mathematical ability at the country level gives .190 / (1.000 + .190) = .160. This indicates that 16.0% of the variance in mathematical ability exists between countries, and 83.6 % of the variance exists within countries.

Step 4: Explaining differences between countries by contextual variables. I added the country-level variables shortage of math teachers and the student–teacher ratio to the between model, and correlated them with mathematical ability. The fit of this model was good: $\chi^2(87) = 98.66$, $p = .18$, RMSEA = .003, CFI = .998. The correlation with student–teacher ratio was negative and statistically significant ($r = -.24$, $p < .05$). This indicates that, as may be expected, countries with a larger student–teacher ratio have lower average scores on mathematical ability. Shortage of math teachers was not significantly related to mathematical ability.

Conclusion. Mathematical ability as measured by these nine items can be considered measurement invariant across countries. There are considerable differences in mathematical ability across countries (16.0% of the variance), part of this variance is explained by country-level differences in the student to teacher ratio.

Discussion

I outlined and illustrated how two-level factor analysis can be used to test for strong factorial invariance and make valid across-country comparisons on latent variables. The two-level model separates country differences in common factors, which are represented by the common factor variance at the between level, from country differences in residual factors, which are represented by residual variances at the between level. Additional country-level variables can be included in the model to explain the differences between countries.

Sources of Measurement Bias

The residual variance at the country level represents the influence of (unknown) factors that bias the measurement by the associated indicator. One large potential source of bias is the translation of items. If an item contains a term that may be ambiguously translated across countries, this will lead to an additional source of variance in this item. Another general source of bias may be country differences in the familiarity of respondents with the subject of an item. For example, an item containing the word "war" may have a different meaning in countries where people actually experienced a war situation than in countries where this is not the case. As outlined, the proposed approach facilitates the inclusion of country-level variables to explain country-level bias in items. Very often, the supposedly biasing factor may not be operationalized. In this case, quantifying the amount of variance caused by other factors than the common factor may already be more informative than simply assuming that the measurements are unbiased. From the simulation research by Jak et al. (2013), it appeared that, when bias is present, accounting for bias at the between level by freely estimating residual variance leads to unbiased estimates of the factor variance. Hence, even if bias is found, one can still correctly evaluate the relation of the country-level common factor with other country level variables of interest, while accounting for the bias by freely estimating the residual variance of biased items.

Problems Regarding the Number of Countries

A problem that is often encountered in MLSEM is that the number of countries is not enough to obtain reliable estimates, or to be able to fit the hypothesized model (Maas & Hox, 2005). These problems arise especially when the between-level model is complex. An important feature of the two-level factor model representing strong factorial invariance is that the model is very restrictive, which facilitates the estimation of the model. Simulation research by Jak et al. showed that with 50 clusters of 25 observations, the parameter estimates of the factor variances were unbiased. With cross-national data, the samples per country are much larger than 25, leading to even more favorable results. Stegmueller (2013) found that MLR with 20 countries with 500 observations each leads to acceptable results if the model is restrictive enough. When strong factorial invariance holds, the only parameter to be estimated at the between level is the factor variance. For each residual variance that needs to be estimated, the estimation becomes increasingly complex. In our examples, it was possible to estimate all residual variances. However, if this leads to problems, one may start with a model with equal factor loadings across levels where all residual variance is fixed at 0, and then test the significance of residual variance by freeing the parameters one at the time for each item.

Advantages and Disadvantages of the Approach

Advantages of the presented procedure are that model fitting is straightforward (see the supplementary material), that the decomposition of the variance at each level can be investigated, that it allows researchers to account for country bias by estimating residual variance, and that it easily allows for the explanation of bias by country-level variables.

A disadvantage of the presented procedure is that both differences in factor loadings and differences in intercepts across countries show up as residual variance. Consequently, when detecting residual variance, it remains unknown whether the bias stems from differences in factor loadings and/or differences in intercepts. Hence, if the goal of the analysis is not an overall test on the presence of measurement bias, but one wants to differentiate between the two types of bias, other methods may be more appropriate. One could, for example, use multigroup analysis as presented by Davidov et al. (2012), random item effects modeling (Verhagen & Fox, 2012), or tests of approximate invariance (Muthén & Asparouhov, 2013). It has to be noted that, for all bias detection methods, it would be a problem if all items would be affected by the same biasing factor to the same degree. That is, strong factorial invariance does not rule out uniform construct-level bias, as explained in the cross-cultural context by Little (2000).

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Weak and strong factorial invariance are also referred to as metric and scalar invariance (e.g., Vandenberg & Lance, 2000), respectively.
2. As suggested by a reviewer, in the terminology of Tay et al., this model could be called "strict metric isomorphism," as it imposes one more restriction in addition to strong metric isomorphism. To avoid introducing new terminology, I will refer to this as the model "representing strong factorial invariance."
3. Sometimes, estimating a saturated between-level model is not possible due to computer memory problems. Especially with categorical data, such as correct/incorrect responses, the computations are a lot heavier than with continuous data (Grilli & Rampichini, 2007). As an alternative, one could fit the factor model to the total data, while correcting the fit statistics and standard errors for the multilevel structure (this option is called "type = complex" in Mplus). Another, more cumbersome, option is to treat the countries as different studies and use meta-analytic structural equation modeling (MASEM) as proposed by Cheung et al. (2008). These authors proposed fixed effects MASEM-analysis. However, at the stage where we just want to find an accurate measurement model without further constraints, random-effects MASEM (Cheung, 2014) may be the preferred approach. Random effects MASEM is less restrictive than fixed effects MASEM, because it allows for heterogeneity across the countries, whereas fixed effects meta-analysis assumes equal population correlation matrices across countries.
4. The exact data sets for this and the next example can be downloaded from the respective websites from ESS and OECD, but are also available from the author upon request.
5. One might argue that treating the responses on a 4-point scale as continuous may lead to incorrect results (Rutkovski & Svetina, 2014). Therefore, ran the same analyses using categorical data analysis to compare the results. However, Mplus was not able to converge to a solution for this model, so continuous data analysis was the only option in this case.
6. As another way to avoid capitalizing on chance, it is recommended to apply a Bonferroni-correction on the alpha level, so that alpha is divided by the number of residual variances under evaluation. Also note that when testing whether a variance is zero one could actually apply one-sided testing as explained in Stoel, Garre, Dolan, & Van de Wittenboer (2006).

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*, 67-91. doi:10.1111/j.1745-3984.1992.tb00368.x

Asparouhov, T., & Muthén, B. (2007). Computationally efficient estimation of multilevel high-dimensional latent variable models. In *Proceedings of the 2007 Joint Statistical Meetings, Section on Statistics in Epidemiology* (pp. 2531-2535). Salt Lake City, Utah.

Barendse, M. T., Oort, F. J., & Garst, G. J. A. (2010). Using restricted factor analysis with latent moderated structures to detect uniform and non-uniform measurement bias: A simulation study. *Advances in Statistical Analysis*, *94*, 117-127. doi:10.1007/s10182-010-0126-1

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238-246.

Billiet, J. (2003). Cross-cultural equivalence with structural equation modeling. In J. A. Harkness, F. J. R. van de Vijver, & P. Ph. Mohler (Eds.), *Cross-cultural survey methods* (pp. 247-264). New York, NY: Wiley.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*, 230-258.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*, 456-466.

Byrne, B. M., & van de Vijver, F. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing*, *10*, 107-132.

Cheung, M. W. -L. (2014). Fixed- and random-effects meta-analytic structural equation modeling: Examples and analyses in R. *Behavior Research Methods*, *46*, 29-40. doi:10.3758/s13428-013-0361-y

Cheung, M. W.- L., Leung, K., & Au, K. (2006). Evaluating multilevel models in cross-cultural research: An illustration with social axioms. *Journal of Cross-Cultural Psychology*, *37*, 522-541.

Davidov, E., Dülmer, H., Schlüter, E., Schmidt, P., & Meuleman, B. (2012). Using a multilevel structural equation modeling approach to explain cross-cultural measurement noninvariance. *Journal of Cross-Cultural Psychology*, *43*, 558-575.

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology*, *40*, 55-75. doi:10.1146/annurev-soc-071913-043137

Dolan, P., Peasgood, T., & White, M. (2008). Do we really know what makes us happy? A review of the economic literature on the factors associated with subjective well-being. *Journal of Economic Psychology*, *29*, 94-122.

ESS Round 6: European Social Survey. (2014). (ESS-6 2012 Documentation Report, Edition 2.1). Bergen: European Social Survey Data Archive, Norwegian Social Science Data Services.

Fischer, R., & Schwartz, S. (2011). Whence differences in value priorities? Individual, cultural, or artifactual sources. *Journal of Cross-Cultural Psychology*, *42*, 1127-1144.

Göbel, S. M., Shaki, S., & Fischer, M. H. (2011). The cultural number line: A review of cultural and linguistic influences on the development of number processing. *Journal of Cross-Cultural Psychology*, *42*, 543-565.

Grilli, L., & Rampichini, C. (2007). Multilevel factor models for ordinal variables. *Structural Equation Modeling*, *14*, 1-25.

Guenole, N. (2016). The importance of isomorphism for conclusions about homology: A Bayesian multilevel structural equation modeling approach with ordinal indicators. *Frontiers in Psychology*, *7*, 289.

Hox, J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Lawrence Erlbaum.

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

Huppert, F. A., Marks, N., Clark, A., Siegrist, J., Stutzer, A., Vittersø, J., & Wahrendorf, M. (2009). Measuring well-being across Europe: Description of the ESS well-being module and preliminary findings. *Social Indicators Research*, *91*, 301-315. doi:10.1007/s11205-008-9346-0

Jak, S. (2014). Testing strong factorial invariance using three-level structural equation modeling. *Frontiers in Psychology*, *5*, 745.

Jak, S., Oort, F. J., & Dolan, C. V. (2013). A test for cluster bias: Detecting violations of measurement invariance across clusters in multilevel data. *Structural Equation Modeling*, *20*, 265-282.

Jak, S., Oort, F. J., & Dolan, C. V. (2014). Measurement bias in multilevel data. *Structural Equation Modeling*, *21*(2), 31-39.

Little, T. D. (2000). On the comparability of constructs in cross-cultural research: A critique of Cheung and Rensvold. *Journal of Cross-Cultural Psychology*, *31*, 213-219.

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*, 86-92.

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods*, *10*, 259-284. doi:10.1037/1082-989X.10.3.259

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

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525-543.

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

Molenaar, D., Dolan, C. V., Wicherts, J. M., & van der Maas, H. L. J. (2010). Modeling differentiation of cognitive abilities within the higher-order factor model using moderated factor analysis. *Intelligence*, *38*, 611-624. doi:10.1016/j.intell.2010.09.002

Muthén, B. O. (1990). *Mean and covariance structure analysis of hierarchical data*. Los Angeles: University of California, Los Angeles.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research*, *22*, 376-398.

Muthén, B. O., & Asparouhov, T. (2013). *New methods for the study of measurement invariance with many groups* [Technical report]. Available from http://www.statmodel.com

Muthén, L. K., & Muthén, B. O. (1998-2015). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author.

Organisation for Economic Co-Operation and Development. (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. Paris, France.

Poortinga, Y. H. (1989). Equivalence of cross-cultural data: An overview of basic issues. *International Journal of Psychology*, *24*, 737-756.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, *69*, 167-190.

Radloff, L. S. (1977). The CES-D Scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*, 385-401.

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement*, *74*, 31-57.

Selig, J. P., Card, N. A., & Little, T. D. (2008). Latent variable structural equation modeling in cross-cultural research: Multigroup and multilevel approaches. In F. J. R. Van de Vijver, D. A. van Hemert, & Y. Poortinga (Eds.), *Multilevel analysis of individuals and cultures* (pp. 93-119). Mahwah, NJ: Lawrence Erlbaum.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structures between groups. *British Journal of Mathematical and Statistical Psychology*, *27*, 229-239.

Stegmueller, D. (2013). How many countries for multilevel modeling? A comparison of frequentist and Bayesian approaches. *American Journal of Political Science*, *57*, 748-761.

Steiger, J. H., & Lind, J. C. (1980, May). *Statistically based tests for the number of common factors*. Paper presented at the annual meeting of the Psychometric Society, *Iowa City, IA*.

Stoel, R. D., Garre, F. G., Dolan, C. V., & van den Wittenboer, G. (2006). On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints. *Psychological Methods*, *11*, 439-455.

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393-408.

Tay, L., Woo, S. E., & Vermunt, J. K. (2014). A conceptual framework of cross-level isomorphism: Psychometric validation of multilevel constructs. *Organizational Research Methods*, *17*, 77-106. doi:10.1177/1094428113517008

van de Vijver, F. J. R., van Hemert, D. A., & Poortinga, Y. H. (2008). Conceptual issues in multilevel models. In F. J. R. Van de Vijver, D. A. Van Hemert, & Y. H. Poortinga (Eds.), *Multilevel analysis of individuals and cultures* (pp. 3-26). New York, NY: Lawrence Erlbaum.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4-70.

Verhagen, A. J., & Fox, J.-P. (2012). Bayesian tests of measurement invariance. *British Journal of Mathematical and Statistical Psychology*, *66*, 383-401.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281-324). Washington, DC: American Psychological Association.

Yuan, K. H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. In M. E. Sobel & M. P. Becker (Eds.), *Sociological methodology* (pp. 165-200). Washington, DC: American Sociological Association.