Using a multilevel structural equation modeling approach to explain cross-cultural measurement noninvariance

Davidov, Eldad; Dülmer, Hermann; Schlüter, Elmar; Schmidt, Peter; Meuleman, Bart

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Journal of Cross-Cultural Psychology
43(4) 558–575
©The Author(s) 2012
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0022022112438397
jccp.sagepub.com
SAGE

Using a Multilevel Structural Equation Modeling Approach to Explain Cross-Cultural Measurement Noninvariance

Eldad Davidov[1], Hermann Dülmer[2], Elmar Schlüter[2], Peter Schmidt[3], and Bart Meuleman[4]

Abstract

Testing for invariance of measurements across groups (such as countries or time points) is essential before meaningful comparisons may be conducted. However, when tested, invariance is often absent. As a result, comparisons across groups are potentially problematic and may be biased. In the current study, we propose utilizing a multilevel structural equation modeling (SEM) approach to provide a framework to explain item bias. We show how variation in a contextual variable may explain noninvariance. For the illustration of the method, we use data from the second round of the European Social Survey (ESS).

Keywords

configural, metric, and scalar invariance, multilevel confirmatory factor analysis (CFA) / multilevel structural equation modeling (SEM), European social survey, comparisons over time and/or countries

When investigating a theory and applying an instrument in different countries or over time, a key concern of researchers is to ensure that the measurement of the relevant constructs is invariant cross-nationally or over time. Testing for invariance of measurements across countries and over time is necessary before meaningful comparisons of relationships and means may be conducted (Billiet, 2003). Horn and McArdle (1992) define *measurement invariance* as "whether or not, under different conditions of observing and studying phenomena, measurement operations yield measures of the same attribute" (p. 117). In other words, invariance guarantees that items are perceived in a similar way and that constructs are represented on the same measurement scale (i.e., with equal factor loadings and intercepts) (see Byrne & van de Vijver, 2010, p. 108). If invariance is absent, observed differences in means or other statistics might reflect differences in

[1]University of Zurich, Zurich, Switzerland
[2]University of Cologne, Cologne, Germany
[3]State Research University Higher School of Economics (HSE), Moscow, Russian Federation
[4]University of Leuven, Leuven, Belgium

Corresponding Author:
Eldad Davidov, Institute of Sociology, University of Zurich, Andreasstrasse 15, Zurich, CH-8050
Email: davidov@soziologie.uzh.ch

systematic biases of response across countries or different understanding of the concept, rather than substantive differences per se (Steenkamp & Baumgartner, 1998). Equally important, findings of no difference between countries do not ensure the absence of "real" differences.

To date, cross-cultural research on invariance has focused mainly on testing for the presence or absence of invariance of theoretical concepts (see, e.g., Ariely & Davidov, 2010; Billiet, 2003; Meuleman, Davidov, & Billiet, 2009; Davidov, 2008, 2009; Davidov, Schmidt, & Schwartz, 2008; De Beuckelaer, Lievens, & Swinnen, 2007; Van der Veld & Saris, 2011). Typically, these tests have been conducted using multiple group confirmatory factor analysis (MGCFA: Bollen, 1989; Jöreskog, 1971, but for other methods see, e.g., Davidov, Schmidt, & Billiet, 2011). Results in many of these studies were able to demonstrate that the assumption that item intercepts (i.e., the expected item score for a respondent with a zero score on the latent variable) are equal across groups is particularly problematic. However, this type of research has largely neglected investigating *why* invariance is absent (for a notable exception, see Byrne & van de Vijver, 2010; for studies tackling a similar question within a multidimensional scaling [MDS] framework, see Fischer, Milfont, & Gouveia, 2011; Fontaine, Poortinga, Delbeke, & Schwartz, 2008). This neglect is unfortunate because findings of noninvariance may reveal meaningful cross-cultural differences.

In the present study we show how multilevel structural equation modeling (SEM) can be used to explain noninvariance. Whereas lower levels (i.e., configural or metric) of invariance are often supported by the data in cross-national studies, this becomes increasingly seldom when higher levels (i.e., scalar) of invariance are tested across cultures or countries. Indeed, scalar noninvariance constitutes one of the most serious threats to cross-cultural research, and it is also the focus of the present study. By using multilevel SEM to explain scalar noninvariance, we are not proposing a new technique, particularly because this technique has been around now for more than two decades (see, e.g., Cheung & Au, 2005; Hox, 2002; Muthén, 1989, 1994). Rather, we show how it may be used to provide a framework to explain item bias across countries. Thus, the application of multilevel SEM for this purpose is new.

The study proceeds as follows. First, we briefly describe the concept of measurement invariance and how it can be tested. Next, we report strategies suggested in the literature to address the problem of noninvariance. In the next step, we specify how multilevel analysis may be used to address and explain noninvariance. Finally, we turn to an empirical example that demonstrates the procedure. We finalize with some conclusions and limitations.

Testing for Measurement Invariance

There can be little doubt that invariance tests have proven themselves as a necessary step in cross-cultural analyses (for a general discussion on invariance tests, see, e.g., Meredith, 1993). In these types of studies, MGCFA is commonly used to conduct the tests (for an overview of different methods to test for invariance, see, e.g., De Beuckelaer, 2005). Here one typically distinguishes between three important levels of invariance: configural, metric, and scalar.

Configural invariance is the lowest level of invariance. It indicates that the same items load on the same latent variables across groups (which may be different countries, cultures, regions, or time points). Configural invariance is supported by the data when a model that specifies which items measure each latent variable fits the data well in all countries. Configural invariance, however, does not yet guarantee that it is measured on the same scale (Steenkamp & Baumgartner, 1998).

A higher level of invariance, metric invariance, assesses a necessary condition for invariance of meaning. Selig, Card, and Little (2008, p. 95) use the term *weak factorial invariance* to describe this level of invariance. Metric invariance indicates that the factor loadings of the indicators are equal. If metric invariance is present, it implies that the latent variable has equal scale intervals over countries. As a result, it allows a meaningful comparison of relationships (unstandardized regression coefficients, covariances) between the latent construct and other concepts across groups (Steenkamp & Baumgartner, 1998). Metric invariance is tested by restricting each factor loading of a corresponding item to be the same across groups.

$$\Lambda^1 = \Lambda^2 = \Lambda^3 = \ldots = \Lambda^G \tag{1}$$

where G = number of groups and Λ = vector of factor loadings.

Metric invariance is supported if such a model fits the data well. Metric invariance must be established for subsequent tests to be meaningful.

Both configural and metric invariance are tested by using information on the covariances between the items. They are not sufficient if the goal of the analysis is to compare means across groups. To justify comparing means, a third, higher level of invariance is necessary, scalar invariance. Scalar invariance additionally requires that the intercepts of each indicator are identical across groups:

$$\tau^1 = \tau^2 = \tau^3 = \ldots = \tau^G, \tag{2}$$

where G = number of groups and τ = vector of item intercepts.

Item intercepts are the expected item scores for respondents that have a zero score on the latent variable. Once the requirement of equal intercepts has been fulfilled, meaningful latent mean comparison of the theoretical concepts becomes possible (Cheung & Rensvold, 2002; De Beuckelaer, 2005; Harkness, van de Vijver, & Mohler, 2003; Hui & Triandis, 1985; Meredith, 1993; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). The equality of intercepts concretely implies that all observed mean differences in the items must be conveyed through mean differences in the latent factor, instead of being a product of cross-country differences in item functioning.

To assess scalar invariance, one thus additionally constrains the intercepts to be equal across groups and tests the fit of the model to the data. As we have mentioned before, this level of invariance is especially seldom achieved, when groups (e.g., countries, but also gender and age groups, cultural groups, or regions) are compared (see, e.g., Steinmetz, Schmidt, Tina-Booh, Wieczorek, & Schwartz, 2009). In sum, a meaningful mean comparison across groups requires three levels of invariance: configural, metric, and scalar. Only if the three levels of invariance are established can meaningful cross-country mean comparisons be carried out. It should be noted, however, that it might become very tedious to use MGCFA to test for invariance when the number of countries or units becomes very large (i.e., more than 20; see Jak, Oort, & Dolan, 2011).

What Can Be Done When Cross-Group Invariance Is Absent?

What can one do when cross-group invariance is absent? The literature provides only a few guidelines offering suggestions for dealing with such a situation. One commonly used strategy when full invariance is absent is to resort to *partial* invariance. Several authors have proposed that two indicators measuring the underlying latent variable with equal loadings and/or intercepts are sufficient to guarantee partial metric and/or scalar invariance (Byrne, Shavelson, & Muthén, 1989; Steenkamp & Baumgartner, 1998; for criticisms, see, e.g., De Beuckelaer & Swinnen, 2011). According to this approach, partial invariance is sufficient for making valid cross-group comparisons (for an application, see Meuleman et al., 2009). When less than two items per latent variable have equal loadings and/or intercepts, these authors suggest that cross-cultural comparisons are biased and therefore problematic. A second approach consists of comparing only a subset of countries (or other groups) where invariance of the involved concepts does hold (Byrne & van de Vijver, 2010). Welkenhuysen-Gybels, van de Vijver, and Cambré (2007), for example, discuss various clustering techniques to detect groups of countries for which constructs are measured in a cross-culturally comparable way. Although helpful in several cases, these two approaches are not entirely satisfactory. The first proposal does not clarify what steps could be additionally undertaken in those cases where even partial invariance is absent. The second approach may drastically reduce the number of cultural groups included in the study. A third approach proposed in the literature is to decrease the number of items and delete those items whose parameters are very different across groups (Welkenhuysen-Gybels, 2003). However, when this approach is applied, one has to address the question of whether the meaning of the concept has changed after the item reduction (Byrne & van de Vijver, 2010). A fourth, more flexible approach was suggested by Muthén (1985, 1989; see also Brown, 2006, pp. 204-206; Lee, Little, & Preacher, 2011; Oort, 1992, 1998). According to this approach, one could use a multiple indicators multiple causes (MIMIC) model to explain item bias. For instance, if a certain item functions differently across categories of some individual characteristic such as gender or age, one could account for this variability by regressing the item on that variable. If the effect of gender or age on the item is significant, it is an indication that the item functions differently across gender or age groups and is thus noninvariant. Jak et al. (2011) indicate that this method is useful to detect scalar noninvariance but is less straightforward to detect metric noninvariance. However, recent developments in latent interaction modeling may provide feasible ways to also detect metric noninvariance using this approach.

When the variance is due to a variable on a higher level of analysis, then we have to account for the different levels of analysis. Thus, we propose a fifth approach to deal with noninvariance. In this approach, one can try to *explain* noninvariance and account for the variance of the items on the contextual level of analysis by introducing contextual predictor variables in a multilevel analysis (Schlüter & Meuleman, 2009). In this respect, it is suggested that noninvariance can be viewed as a useful source of information on cross-group differences (e.g., Medina, Smith, & Long, 2009; Poortinga, 1989; Schlüter & Meuleman, 2009). Although it has already been referred to by some authors (see, e.g., Hox, de Leeuw, & Brinkhuis, 2010; Jak et al., 2011) and although the technique is not new (see, e.g., Cheung & Au, 2005; Hox, 2002; Muthén, 1989, 1994), to the best of our knowledge this possibility has not yet been explicated and systematically applied for the goal of explaining measurement noninvariance across contextual units of analysis such as countries or cultures. Its distinct advantage compared to the other approaches is that it can potentially *explain* noninvariance in a substantive way. If the context level is represented by countries, for instance, this approach uses country information as a possible source of bias to explain differences in items that display large cross-country differences. Finding the source of bias can deliver useful information as to how certain scales may be improved for cross-cultural research. Its main difference from the fourth approach is that contextual-level rather than individual-level information is used to explain item bias.

Using Multilevel Techniques to Explain Measurement Noninvariance

Multilevel structural equation modeling (MLSEM) has been known for more than two decades (cf., Cheung & Au, 2005; Hox, 2002; Muthén, 1985, 1994). However, only after its inclusion in structural equation modeling computer programs like Mplus (Muthén & Muthén, 1998-2010) in recent years has its application become more accessible to applied researchers. Similar to multilevel regression models, MLSEM decomposes the variability of the indicators into individual ("within") and contextual ("between"; e.g., country) variability.

The procedure of using MLSEM techniques to explain noninvariance includes two steps. In the first step, a multilevel confirmatory factor analysis (CFA) is conducted. In a multilevel CFA, we account for variations in the indicators both across individuals and across contexts by individual- and contextual-level latent variables. Figure 1 illustrates a two-level CFA with one latent factor

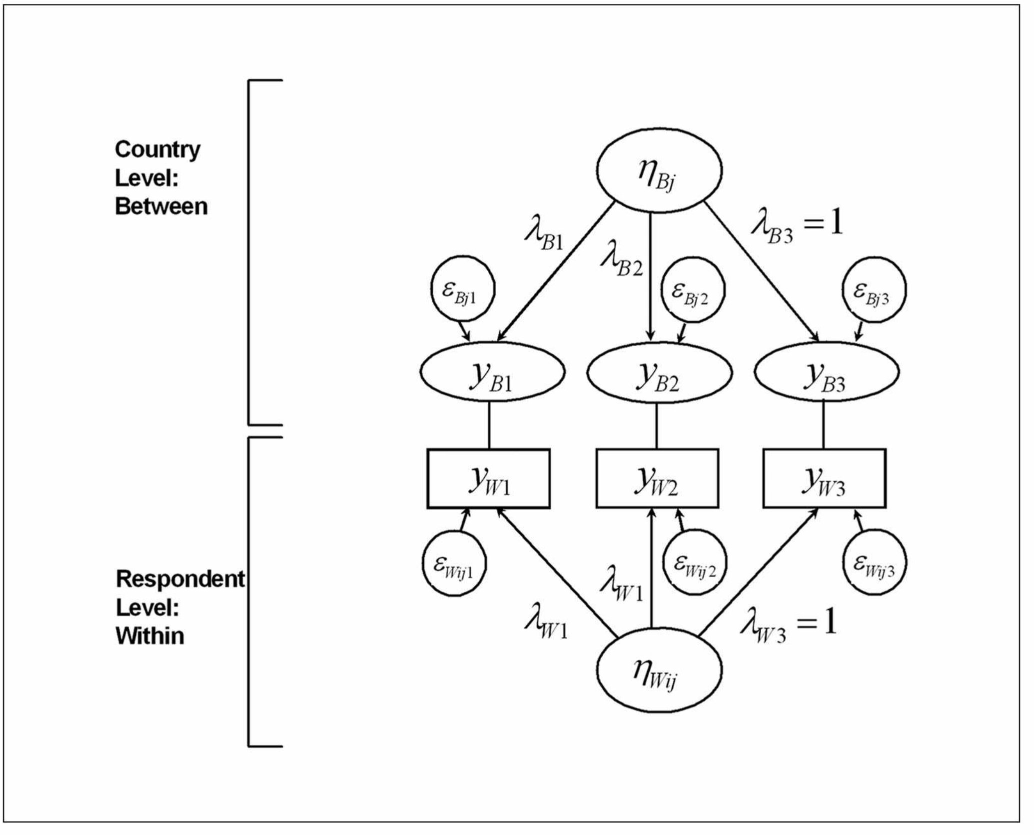

Figure 1. A Two-Level CFA With Three Indicators
Note: Rectangles represent $k = 3$ indicators on the within-level; one-sided arrows represent causal effects; the large circles of η_{Wij} and η_{Bj} represent the latent variable on the within and between levels, respectively; the small circles next to the rectangles refer to the within-level error term ε_W for respondent *i* of country *j* on indicator variable *k*; the large circles of *y* on the between level refer to the indicator variable on the between level; the small circles next to the indicators on the between level refer to the between-level error term ε_B (usually called random term in multilevel analysis).

at Level 1 (within) and one latent factor at Level 2 (between) with k = 3 Level 1 indicator variables.

The two-level CFA model can be written as follows (cf., also Muthén, 1991, p. 344):

Level 1 (Within):	Level 2 (Between):	
$y_{ijk} = \alpha_{jk} + \lambda_{Wk} \cdot \eta_{Wij} + \varepsilon_{Wijk}$	$\alpha_{jk} = \upsilon_k + \lambda_{Bk} \cdot \eta_{Bj} + \varepsilon_{Bjk}$	(3)

where

- y_{ijk} refers to the observed value of respondent *i* of country *j* on indicator variable *k*,
- α_{jk} refers to the intercept of indicator variable *k* in country *j*,
- υ_k refers to the cross-country grand intercept of indicator variable *k* (i.e., the grand mean when the between-level latent variable equals zero),
- η_{Wij} refers to the score of respondent *i* of country *j* on the within-level latent η_W,

- η_{Bj} refers to the score of country j on the between-level latent variable η_B,
- λ_{Wk} refers to the within-level factor loading λ_W of indicator variable k,
- λ_{WB} refers to the between-level factor loading λ_B of indicator variable k,
- ε_{Wijk} refers to the within-level error term ε_W for respondent i of country j on indicator variable k, and
- ε_{Bjk} refers to the between-level error term ε_B (usually called random intercept term in multilevel analysis) for country j on indicator variable k.

The within part of Equation (3) and the between part of the equation are connected in a multilevel CFA via the intercept α_{jk} of country j on indicator k: The country-specific item intercepts α_{jk} for indicator k on the within part are at the same time the dependent variable in the between part equation. This connection is depictured in Figure 1 by a straight line between the within- and between-level components of the indicators. Each country j's indicator intercept—α_{jk}—is random at the between level (country level). The variability of the country-specific intercepts α_{jk} of an indicator variable k is explained in the between-level by the latent variable η_{Bj}. The non-explained variability in the countries' intercepts α_{jk} after controlling for the effect of the between-level latent variable is captured by the country error term ε_{Bjk}.

A close connection exists between this two-level CFA model and the measurement invariance framework sketched above (see Fontaine, 2008, for a more systematic elaboration of this point). Measurement noninvariance can appear in various ways in two-level CFA. Unequal factor loadings across groups can be modeled by allowing one or more random slopes for the within-level factor loadings (Schlüter & Meuleman, 2009). Cross-group intercept differences (deviations from scalar invariance) show up in the between-level error terms ε_{Bjk}. Concretely, nonzero error terms indicate that the country means for some items are not equal to what is expected based on the between-level latent mean. In other words, substantial between-level error variance in the indicators points in the direction of unequal item intercepts or deviations from scalar equivalence. The connection between MLSEM and measurement invariance is also clear from the fact that several authors have argued that to perform meaningful MLSEM, certain assumptions are made about measurement invariance. Cheung, Leung, and Au (2006, p. 523), for example, stress that the within-factor structure should be the same across groups and propose to test this assumption by using meta-analytic structural equation modeling (MASEM). Fontaine (2008, pp. 77-78) similarly stresses that relations between latent factors and indicators should be identical (or very similar) across groups and that the country-level error terms should be (very close to) zero.[1] In this study, we take the position that drawing meaningful conclusions from MLSEM presupposes equal factor loadings and item intercepts.

When these assumptions are not met, correcting for the measurement noninvariance is a sensible option (Fontaine, 2008, p. 78). This is done in the second step of the procedure we propose: accounting for cross-group differences in the parameters (such as intercepts) by including individual and/or contextual predictors in the model (see Jak et al., 2011). In this step, the multilevel CFA (cf., Hox, 2002; Muthén, 1994) is extended to a multilevel SEM (cf., Muthén, 1994; Selig et al., 2008), which allows the explanation of measurement noninvariance by individual and/or contextual variables. This approach is not an alternative to the cross-cultural comparison of the theoretical concepts of interest. Instead, it constitutes a useful test to explain *why* invariance does not hold.

In this step, we include contextual predictors in order to further explain Level 2 variability of the indicators (α_{jk}). By means of these contextual predictors, we try to reduce the unexplained country-level variance of the indicators (ε_{Bjk}). If the remaining variability in the intercept was fully explained, then the between-level error term ε_{Bjk} should become zero, and measurement noninvariance is fully accounted for. Assuming that the context is the country, then country characteristics that are included as predictors in Level 2 could be aggregates of individual-level variables such as employment status or education, or variables that characterize the country level such as the level of human development in a country, policies, history, or economic conditions.

In the following we will illustrate, with a simple example using data from the European Social Survey (ESS), how the method may be used to explain scalar noninvariance of one of the indicators measuring the value universalism from the value theory of Schwartz (1992). Previous studies have demonstrated that the value measurements in the ESS fail to display scalar invariance (Davidov, 2008; Davidov et al., 2008). The present application will show how using even one contextual variable may be very fruitful in explaining noninvariance. In this case of one contextual variable only, the model would be equivalent to the use of a MIMIC multigroup model with *n* groups (see Brown, 2006, pp. 204-206).

Empirical Illustration

Theoretical Considerations

Schwartz (1992) proposes 10 basic universal human value types, each with distinct motivational emphases. In the present example, we focus on the value type universalism because it is the only value that is measured by three indicators (all other values in the theory are measured by only two questions each). The theory suggests at least three main elements for universalism (although later developments have further extended the dimensions of this value). The first is related to the importance of equal treatment and equal opportunities for everyone. The second element taps the importance of protecting the environment. The third is related to broad-mindedness and tolerance. These elements are considered to be closely linked with each other (Schwartz, 1994). Although the theory postulates that this value and its three elements should be found universally, its level and the way it is understood may differ across cultures.

Inglehart (1997, pp. 9, 14-15, 67) proposed that cross-country variations in the level and understanding of values may be accounted for by country differences in economic and technological development. There are two key hypotheses in Inglehart's (1990, 1997) approach. The first asserts that "one places the greatest subjective value on things that are in relatively short supply" (the scarcity argument, see Inglehart, 1997, p. 33). The second suggests that "one's basic values reflect the conditions that prevailed during one's pre-adult years" (the socialization argument, see Inglehart, 1997, p. 33). Based on Maslow's (1954) need hierarchy, these two assumptions led Inglehart to expect an intergenerational individual value change from more fundamental materialist value priorities (physical and economical security) to higher order postmaterialist value priorities (belonging, self-expression) in advanced industrial societies (see also Inglehart, 1997, p. 33). This individual-level change is the foundation (Coleman, 1994, p. 8) for a broader societal-level syndrome of postmodernization (Inglehart, 1997). Postmodern societies value, according to Inglehart, greater tolerance for ethnic, cultural, and sexual diversity and place an increasing emphasis on protection of environment, all of which are aspects of universalism. Thus, in our first hypothesis we expect higher scores on the value of universalism in postmodern, advanced industrial countries than in less developed, modern countries (Hypothesis 1). However, Inglehart (1997, p. 242) also states that in less economically advanced societies where air and water pollution are far worse than in advanced industrial societies, environmental protection is less a postmodern concern for quality of life **but rather** a matter of physical health. The latter concern, however, gradually fades in advanced industrial societies. This individual-level expectation is the foundation for our second societal-level hypothesis (Coleman, 1994), where we state that environmental protection is expected to be perceived as more important in less developed countries than in postmodern, advanced industrial countries (Hypothesis 2). These considerations explain why the environment item might operate differently depending on a society's developmental level.

Data and Operationalization

The European Social Survey (ESS) includes three questions from the Portrait Values Questionnaire (PVQ; cf., Schwartz et al., 2001) to measure universalism. The questions (gender matched to the respondent) describe a fictitious person, and the respondent is asked to rate the extent to which this person is or is *not* like him or her. The first question (*equality*) is: "He thinks it is important that every person in the world be treated equally. He believes everyone should have equal opportunities in life." The second question (*tolerance* and *understanding*) is: "It is important to him to listen to people who are different from him. Even when he disagrees with them, he still wants to understand them." The third question (*environment*) is: "He strongly believes that people should care for nature. Looking after the environment is important to him." For ease of interpretation, the original scale has been reversed. The reversed scale ranges from 0 (*not like me at all*) to 5 (*very much like me*).

Data were collected in 25 countries that participated in Round 2 of the ESS. The fieldwork of most of these countries was carried out in 2004 and 2005. East and West Germany were treated as separate countries, so that the number of groups in the analysis is actually 26[2] (for a detailed report on data collection and documentation in the participating countries, see www.europeansocialsurvey.org; data may be downloaded at http://ess.nsd.uib.no/).

To measure a country's level of economic development, we use the *Human Development Index* (HDI; cf., United Nations Development Program [UNDP], 2006). This index is also provided in Appendix 1 for each country. In our view, this index best describes how advanced a country is as it combines several criteria, such as a country's standard of living (GDP per capita in purchasing power parity US dollars), the average level of educational attainment, and the country's level of longevity (life expectancy at birth; cf., UNDP 2006, pp. 263 and 276).

Statistical Analyses

We started the analysis by performing a MGCFA and covariance structure analysis (MACS; Sörbom, 1974, 1978) for the universalism value across countries. These techniques allow testing for metric and scalar invariance of the universalism latent variable across countries. As we argued above, this step is required before meaningful comparisons of correlates and means can be conducted (see also Davidov, 2008; Davidov et al., 2008). Next, we conducted multilevel CFA followed by multilevel SEM. In the multilevel CFA, we included one individual-level factor as well as one country-level factor to account for the variability of the universalism indicators on both levels. In the next step, the multilevel SEM, we tried to explain noninvariance of the environment indicator intercept by regressing this indicator and the universalism latent variable (on the between-country level) on the HDI 2004 country-level variable (while accounting for the individual-level universalism latent variable in the model). The software package Mplus version 6.0 (Muthén & Muthén, 1998-2010) was used for the analysis.

Descriptive statistics. First, we observed the correlations and covariances of the indicator variables. Indicators that are supposed to reflect a certain latent variable should correlate highly among each other (Byrne, 2001). Table 1 reports the within- as well as between-level correlations and covariances between the indicators for the simultaneously estimated two-level model. These coefficients are decomposed into their within- and between-countries part. The correlations for the within part of the two-level model range between 0.312 and 0.332. The correlations

Table 1. Correlations, Variances, and Covariances for the Indicators of Universalism

		Within and Between Countries Correlations and Covariances		
		1	2	3
	Within			
1	Equality (ipeqopt)	1.037	*0.332*	*0.312*
2	Underst. Diff. People (ipudrst)	0.357	1.117	*0.321*
3	Environment (impenv)	0.321	0.343	1.019
	Between			
1	Equality (ipeqopt)	0.038	*0.591*	*0.547*
2	Underst. Diff. People (ipudrst)	0.023	0.040	*0.477*
3	Environment (impenv)	0.024	0.021	0.049

Source: ESS data 2004-5.
Note: Italic entries in the upper diagonal are the correlations, entries in the diagonal are variances, and entries in the lower diagonal are covariances; the total sample includes 43,779 respondents from 25 countries (with two German samples: East and West).

for the between part of the latter model are somewhat stronger, ranging from 0.547 to 0.591. All correlations are of a sufficient size, thus enabling us to conduct a CFA for the three indicator variables on both levels.

Testing for invariance. Second, before turning to the multilevel CFA, we started with a multiple group CFA (MGCFA) to evaluate the invariance properties of the universalism variable. We tested for metric and scalar invariance across 26 groups (25 countries). We did not test for configural invariance because with only three indicators the model is just identified. However, previous studies have demonstrated that values display at least configural invariance with the ESS data (Davidov et al., 2008). For the metric invariance model, we constrained the factor loadings between the indicators and the constructs in the model to be the same in all of the countries. If the factor loadings are invariant, we can conclude that the meaning of the universalism value, as measured by the indicators in the ESS, may be identical across all countries, thus allowing covariances or unstandardized regression coefficients to be compared across countries. Although the chi-square statistic is strongly significant ($\chi^2 = 193$, $df = 50$, p value < .0001), various alternative fit indices indicated a good fit between the model and the data that is satisfactory for not rejecting the metric invariance model according to Hu and Bentler (1999) and Marsh, Hau, and Wen (2004) (the comparative fit index, CFI = 0.993; the Tucker-Lewis coefficient, TLI = 0.989; root mean square error of approximation, RMSEA = 0.006; PCLOSE[3] = 1.00; the standardized root mean square residual, SRMR = 0.013). Hence, the metric invariance of the universalism factor model cannot be rejected.

The next step of the MGCFA tested for scalar invariance, a necessary condition for comparing the mean of universalism across countries. This step of MGCFA is augmented with mean structure information (see Sörbom, 1974, 1978). This type of MGCFA is often referred to in the literature as mean and covariance structure (MACS) analysis. It constrains the intercepts of the indicators in the model, in addition to the factor loadings between the indicators and the construct, to be the same in all of the countries. If the factor loadings and the intercepts are invariant, one can legitimately compare value means. The fit indices for the scalar invariance model suggested the rejection of this model ($\chi^2 = 2176$, $df = 100$, CFI = 0.838, TLI = 0.874, RMSEA = 0.021, PCLOSE = 1.00, SRMR = 0.001). Although the RMSEA and SRMR were acceptable according to Hu and Bentler (1999) and Marsh et al. (2004), the decrease in CFI and TLI was too large according to the fit criteria suggested by Chen (2007), leading us to conclude that the scale does not meet the requirements of scalar invariance. For evaluating the fit of the scalar invariance model, we rely on the studies of Cheung and Rensvold (2002) and Chen (2007). Chen (2007) suggested cut-off criteria for *differences* in the global fit measures between the metric and the scalar invariance model. Deterioration in the global fit which is beyond the recommended criteria leads to the rejection of the model.[4]

Next, we considered the modification indices suggested by the program for the full scalar invariance model to detect which cross-country equality constraints on the indicator intercepts were violated by the data. The *modification* index is a lower bound estimate of the expected chi-square decrease that would result when a particular parameter is left unconstrained (Saris, Satorra, & Sörbom, 1987). These modification indices were especially pronounced for the item "environment." In other words, the intercept of the item measuring the importance of the environment displayed the largest cross-country differences, whereas the intercepts of the other two items could be set equal. Thus, in the next sections we will modify the MGCFA model into a two-level CFA and introduce a contextual variable, HDI, to predict the variability that was found in the intercept of environment. Since there was no substantial variability in the factor loadings across countries, we will consider them to be equal.

*Multilevel CFA and multilevel SEM.*In this analysis we first modeled the within and between variability of the universalism indicators in a multilevel CFA model. In the second step we regressed the latent variable of universalism on the between level and the *environment* item on the country-level variable HDI. Thus, we allowed country-level differences in the latent variable and in environment to be predicted by a country-level variable. Table 2 and Figures 2a and 2b contain the results of our multilevel CFA and multilevel SEM analysis without and with the HDI predictor, respectively. The global fit measures of both models presented in the table display a satisfactory model fit.

The empirical results of Model 2, which are depicted in Figure 2b, confirm Hypothesis 1: The higher a country's level of human development (HDI), the more important is the value of universalism for its citizens ($b = 1.165$, $z = 1.871$). Tested one sided, the effect is significant at the 5% level. Thus, respondents in more developed countries score higher on universalism. The empirical results of the model also confirm Hypothesis 2: Environmental protection is significantly less important for people living in advanced industrial countries with a higher HDI than for people living in less developed countries with a lower HDI ($b = -2.965$, $z = -3.757$).[4] Thus, a country's HDI contributes significantly to explain why scalar invariance was not evidenced in the MGCFA. Furthermore, by regressing the item "environment" on HDI on the between level, the residual variance (random component) of that indicator on the between level became insignificant. Hence, country differences in the intercept of "environment" can be traced back completely to differences in the level of human development between the countries.

Discussion and Conclusions

The main *methodological* purpose of this contribution was to explain and illustrate how measurement noninvariance evidenced by MGCFA can be explained by using multilevel SEM. Differences in the intercept of the indicator variables of a latent factor can be modeled in multilevel CFA by including a between-level latent variable and an indicator-specific random term. The variance of this random term can be reduced in a multilevel SEM by regressing the between-level indicator on exogenous between-level variables. Although multilevel CFA/SEM offer a number of further possibilities, we restricted our analyses to explaining noninvariance in

Table 2. Multilevel CFA and Multilevel SEM for Universalism

	Model 1: Two-Level CFA		Model 2 (Including HDI 2004)	
N (Level 2)	25 countries (26 groups)		25 countries (26 groups)	
N (Level 1)	43,779 respondents		43,779 respondents	
AIC	368,050.207		368,042.483	
BIC	368,171.824		368,181.474	
Sample size adjusted BIC	368,127.332		368,130.625	
SRMR within	0.000		0.000	
SRMR between	0.062		0.045	
RMSEA	0.003		0.000	
	b	z	b	z
Factor loadings (Level 2)				
Equality (ipeqopt)	1.000	—	1.000	—
Underst. Diff. People (ipudrst)	0.608	3.666**	0.921	3.197**
Environment (impenv)	0.625	3.277**	1.747	4.599**
Factor loadings (Level 1)				
Equality (ipeqopt)	1.000	—	1.000	—
Underst. Diff. People (ipudrst)	1.069	57.275**	1.069	57.275**
Environment (impenv)	0.960	58.203**	0.960	58.202**
	b	z	b	z
Regression				
Predictor for environment (impenv)				
HDI 2004			–2.965	–3.757**
Predictors for universalism (betw.)				
HDI 2004			1.165	1.871*
Variance	Variance	z		
Latent factor universalism (betw.)	0.038	3.542**		
Latent factor universalism (within)	0.334	42.894**		
Variance Components/Residual Var. Level 2			Variance	z
Universalism (betw.)			0.015	1.943*
Universalism (within)			0.334	42.894**

Note: b = unstandardized regression coefficient. Estimator: Full Maximum Likelihood (ML). Estimates for Level 2 parameters are indented to the right in the first column. Variances/residuals tested one-tailed. Since we formulated hypotheses for the impact of the HDI on environment and universalism (between), the significance level of both *b* coefficients are based on a one-tailed test. AIC = the Akaike information criterion; BIC = the Bayesian information criterion; RMSEA = root mean square error of approximation; SRMR = the standardized root mean square residual. Since multilevel data have a different sample size on different levels, the interpretation of the AIC is more straightforward than that of the BIC and, therefore, the recommended choice (Hox, 2002, p. 46).
$^{*}p \le 0.05$, $^{**}p \le 0.01$.

the *indicator intercept.* Indeed, many researchers are frequently confronted with the situation of scalar noninvariance (where indicator intercepts vary considerably across countries). When indicator intercepts are not similar across countries, mean comparisons of the theoretical constructs of interest are problematic (Billiet, 2003). This approach has the advantage that it may provide an explanation for the absence of invariance. Explanations for noninvariance can follow theory-driven hypotheses, and noninvariance is used as a useful source of information for cross-country differences. Multilevel SEM is a practical method of analysis in this case as it offers researchers the possibility to learn *why* invariance is absent. Although the technique is not new, to the best of our knowledge it has not yet been applied to explain noninvariance in a systematic and theoretically driven way.

We illustrated its use with data from the second round of the ESS and proposed a possible explanation as to why the indicator "environment," one of the indicator variables of Schwartz's universalism value, is scalar noninvariant at the cross-country level of analysis. In addition to

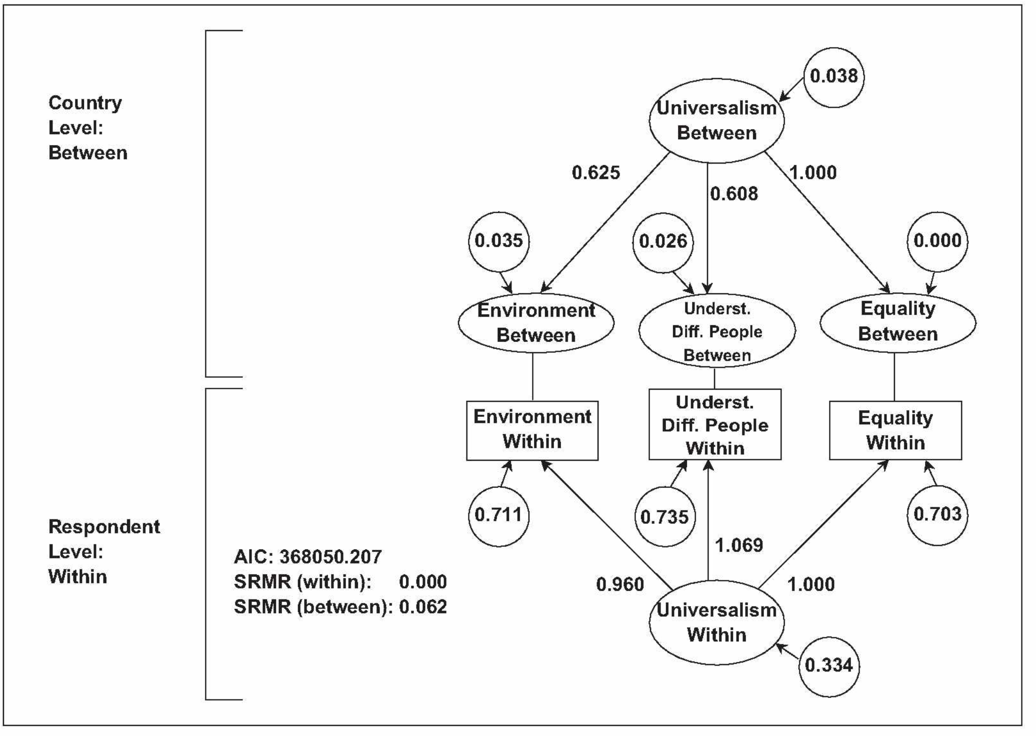

Figure 2a. A Multilevel CFA for Universalism (Model 1)
Note: The residual variance of environment turned out to be insignificant in Model 2 and has been fixed to zero for that reason; for explanations of the components in this figure, see Figure 1; the small circles next to the latent variable universalism in Figure 2a refer to its variance on the within and between levels, respectively; the small circle next to the latent variable universalism on the between level in Figure 2b refers to its prediction error variance; since we formulated hypotheses for the impact of the HDI on environment and universalism (on the between level), the significance level of both *b* coefficients is based on a one-sided test. AIC = the Akaike information criterion; SRMR = the standardized root mean square residual.

this we also tried to explain cross-country differences in the between-level latent factor of universalism: Not regressing the between-level universalism latent variable on HDI would have implied a theoretical and empirical misspecification in this example.[5] We found that a country's level of human development (HDI) successfully explains *why* the intercept of "environment" turned out to be noninvariant in our MGCFA analysis. A country-level economic and technical development as measured by the HDI also contributes significantly to explain differences in the country-level latent variable of Schwartz's universalism across countries. Thus, using multilevel SEM, both of our hypotheses were confirmed. The findings may seem at first counterintuitive from an "Inglehartian" perspective. However, considering the difference between the general concept of universalism and the concept of importance of environment as *one aspect* of universalism makes clear that both hypotheses and findings are in line with Inglehart's reasoning. In less developed countries, both materialists and postmaterialists are more likely to support improved environmental protection (cf., Inglehart, 1997, p. 242).

Because of the limited number of countries included in the analysis, we had to keep the number of contextual explanations to a minimum. Our choice of the HDI variable as a possible cause for variations in the environment indicator was theoretically driven and does not exclude further and/or alternative possible explanations. However, the fact that the residual variance (random

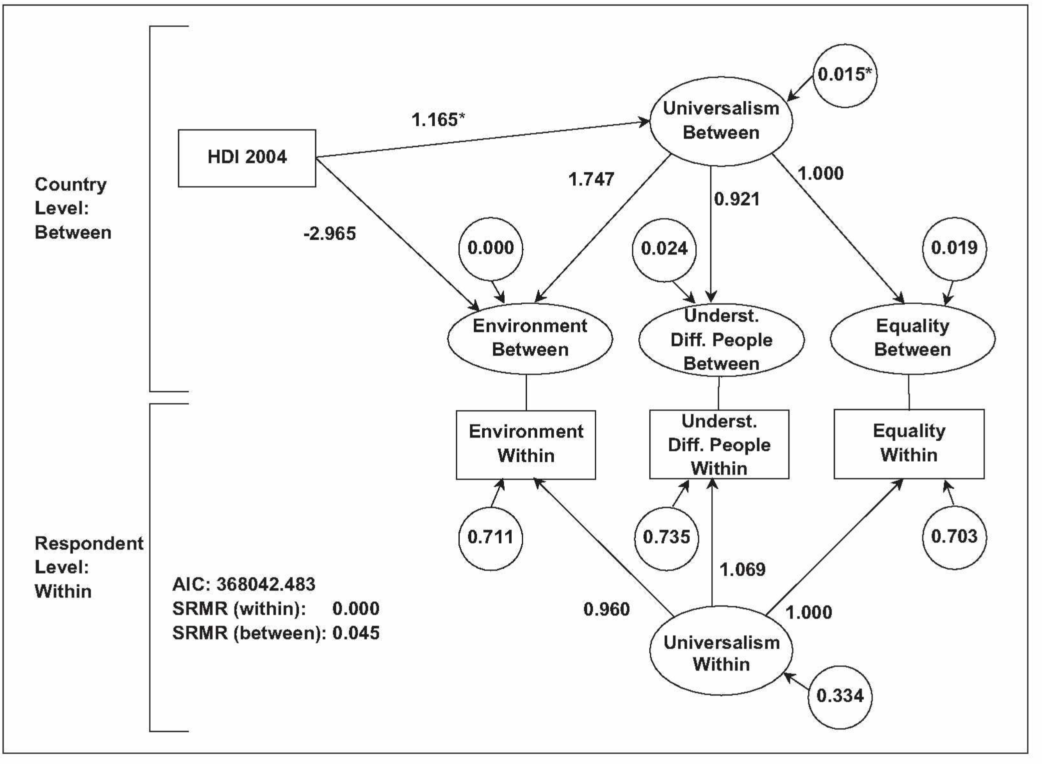

Figure 2b. A Multilevel SEM for Universalism (Model 2)

Note: The residual variance of environment turned out to be insignificant in Model 2 and has been fixed to zero for that reason; for explanations of the components in this figure, see Figure 1; the small circles next to the latent variable universalism in Figure 2a refer to its variance on the within and between levels, respectively; the small circle next to the latent variable universalism on the between level in Figure 2b refers to its prediction error variance; since we formulated hypotheses for the impact of the HDI on environment and universalism (on the between level), the significance level of both *b* coefficients is based on a one-sided test. AIC = the Akaike information criterion; SRMR = the standardized root mean square residual.

$^{*}p < 0.05$

component) of that indicator became insignificant after introducing the HDI variable as a predictor in the multilevel SEM supports the idea that it plays an important role in the explanation of the failure to detect full scalar invariance for that indicator. Future analyses that include a larger set of countries or analyses with a large set of regional units of analysis could account for various macro-level explanations of noninvariance. Finally, although we focused in the illustration on the universalism value, the approach may be applied to other values or other constructs as well. In spite of these limitations, in our point of view, accounting for both contextual-level and individual-level predictors of indicators that fail to display scalar invariance is a promising strategy that offers the possibility to conduct cross-cultural research when invariance cannot be established. Noninvariance then becomes a useful source of information on cross-country differences rather than a hurdle for conducting meaningful cross-country comparative research.

All in all, we hope that our contribution encourages researchers working in the field of cross-cultural research to not refrain from international comparisons when a multiple group CFA fails to establish invariance. Instead, in such cases, a useful strategy could be to look for a theoretical explanation of why invariance does not exist in the first place and to test it. In this respect, multilevel SEM, as an established data analysis method, offers us a powerful new tool.

Appendix

The Level of HDI in 2004 for the Countries in the Analysis[a]

Austria (0.944), Belgium (0.945), Czech Republic (0.885), Denmark (0.943), Estonia (0.858), Finland (0.947), France (0.942), Germany (East and West included separately into our analyses, 0.932, only a common value for both parts of Germany is available), Greece (0.921), Hungary (0.869), Iceland (0.960), Ireland (0.956), Luxembourg (0.945), the Netherlands (0.947), Norway (0.965), Poland (0.862), Portugal (0.904), Slovakia (0.856), Slovenia (0.910), Spain (0.938), Sweden (0.951), Switzerland (0.947), Turkey (0.757), Ukraine (0.774), United Kingdom (0.940)

a. cf., United Nations Development Program (2006).

Acknowledgment

Many thanks to Lisa Trierweiler for the English proofreading of the manuscript.

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) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The work of Peter Schmidt was supported by the Basic research program of the State Research University Higher School of Economics (HSE) Moscow (International Laboratory for Socio-Cultural Research). The other authors received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Although the arguments of Cheung et al. (2006) and Fontaine (2008) bear resemblance to each other, they are not identical. The homogeneity of correlation matrices Cheung et al. (2006) discuss not only implies equal factor loadings across groups, but also presupposes that error covariances and factor (co) variances are similar. The argument developed by Fontaine (2008), on the other hand, implies that, besides factor loadings, item intercepts are also (almost) identical across countries, and thus takes the mean structure of the data into account.
2. Previous work based on simulation studies has shown that performing MLSEM with as little as 26 groups could lead to inaccurate estimation (Meuleman & Billiet, 2009). However, recent simulation studies suggest that Bayesian estimation produces unbiased multilevel estimates, even with group sample sizes as low as 20 (Hox, van de Schoot, & Matthijsse, 2011; Stegmueller, 2011). As a robustness check, all MLSEM models presented in this article were re-estimated using the Bayesian estimation procedure implemented in Mplus 6.0. This led to essentially identical results, strengthening confidence in the validity and reliability of the results. Since we made use of noninformative priors (i.e., the default option in Mplus 6.0), the Bayesian estimates are expected not to be influenced substantially by the choice for certain priors. By means of a simulation study, Hox et al. (2011) have indeed shown that the default estimation procedure in Mplus produces unbiased estimates for a model very similar to the models estimated here.
3. PCLOSE (or the so-called probability of close fit) is a one-sided test of the hypothesis that RMSEA is not larger than .05, the alternative hypothesis being that RMSEA is larger than .05. Values of PCLOSE close to one are indicative of close-fitting models.

4. Saris, Satorra, and van der Veld (2009) have demonstrated that this test of invariance, although very popular, may be too strict. They instead proposed to consider the power of the test and the expected parameter change information. However, applying their approach is beyond the scope of our present study.
5. Since the included countries were not randomly sampled from Europe, we will use the *z* values exclusively in a descriptive sense as a pragmatic criterion to distinguish empirically significant from empirically insignificant effects. Rerunning the model using the Bayesian estimation procedure in Mplus (Muthén & Muthén, 1998-2010) produced essentially the same results.
6. If there are no theoretical reasons to regress the between-level latent variable on the between-level exogenous predictor, it is also possible to allow them to covary (see Jak et al., 2011).

References

Ariely, G., & Davidov, E. (2010). Can we rate public support for democracy in a comparable way? Cross-national equivalence of democratic attitudes in the World Value Survey. *Social Indicators Research, 104*(2), 271-286.

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

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

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

Byrne, B. M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. Mahwah, NJ: Erlbaum.

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.

Chen, F. F. (2007). Sensitivity of goodness of fit indices to lack of measurement invariance. *Structural Equation Modeling, 14*, 464-504.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233-255.

Cheung, M. W.- L., & Au, K. (2005). Applications of multilevel structural equation modeling to cross-cultural research. *Structural Equation Modeling, 12*(4), 598-619.

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*(5), 522-541.

Coleman, J. S. (1994). *Foundations of social theory*. Cambridge, MA: Harvard University Press.

Davidov, E. (2008). A cross-country and cross-time comparison of the human values measurements with the second round of the European Social Survey. *Survey Research Methods, 2*(1), 33-46.

Davidov, E. (2009). Measurement equivalence of nationalism and constructive patriotism in the ISSP: 34 countries in a comparative perspective. *Political Analysis, 17*(1), 64-82.

Davidov, E., Schmidt, P., & Billiet, J. (2011). *Cross-cultural analysis: Methods and applications*. New York: Routledge.

Davidov, E., Schmidt, P., & Schwartz, S. H. (2008). Bringing values back in. The adequacy of the European Social Survey to measure values in 20 countries. *Public Opinion Quarterly, 72*(3), 420-445.

De Beuckelaer, A. (2005). *Measurement invariance issues in international management research*. Unpublished dissertation, Limburgs University Centrum, Centrum, Limburg, the Netherlands.

De Beuckelaer, A., Lievens, F., & Swinnen, G. (2007). Measurement equivalence in the conduct of a global organizational survey across six cultural regions. *Journal of Occupational and Organizational Psychology, 80*, 575-600.

De Beuckelaer, A., & Swinnen, G. (2011). Biased latent variable mean comparisons due to measurement noninvariance: A simulation study. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-cultural research: Methods and applications* (pp. 117-147). New York: Routledge.

Fischer, R., Milfont, T. L., & Gouveia, V. V. (2011). Does social context affect value structures? Testing the within-country stability of value structures with a functional theory of values. *Journal of Cross-Cultural Psychology*, *42*(2), 253-270.

Fontaine, J. R. J. (2008). Traditional and multilevel approaches in cross-cultural research: An integration of methodological frameworks. In F. J. R. van de Vijver, D. A. van Hemert, & Y. H. Poortinga (Eds.), *Multilevel analysis of individuals and cultures* (pp. 65-92). New York: Lawrence Erlbaum Associates.

Fontaine, J. R. J., Poortinga, Y. H., Delbeke, L., & Schwartz, S. (2008). Structural equivalence of the values domain across cultures: Distinguishing sampling fluctuations from meaningful variation. *Journal of Cross-Cultural Psychology*, *39*(4), 345-365.

Harkness, J. A., van de Vijver, F. J. R., & Mohler, P. P. (Eds.). (2003). *Cross-cultural survey methods*. New York: John Wiley.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*, 117-144.

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

Hox, J. J., de Leeuw, E. D., & Brinkhuis, M. J. (2010). Analysis models for comparative surveys. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. P. Mohler, B.- E. Pennell, & T. W. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 395-418). Hoboken, NJ: John Wiley.

Hox, J. J., van de Schoot, R., & Matthijsse, S. (2011). *How few countries will do? Comparative survey analysis from a Bayesian perspective*. Unpublished manuscript.

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.

Hui, C. H., & Triandis, H. C. (1985). Measurement in cross-cultural psychology: A review and comparison of strategies. *Journal of Cross-Cultural Psychology*, *16*, 131-152.

Inglehart, R. (1990). *Culture shift in advanced industrial society*. Princeton, NJ: Princeton University Press.

Inglehart, R. (1997). *Modernization and postmodernization: Cultural, economic, and political change in 43 societies*. Princeton, NJ: Princeton University Press.

Jak, S., Oort, F. J., & Dolan, C. V. (2011, March). *A stepwise approach for the detection of measurement bias in multilevel data*. Paper presented at the Structural Equation Modeling Working Group, Marburg, Germany.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*, 409-426.

Lee, J., Little, T. D., & Preacher, K. J. (2011). Methodological issues in using structural equation models for testing differential item functioning. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-cultural research: Methods and applications* (pp. 55-84). New York: Routledge.

Marsh, H. W., Hau, K.- T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, *11*, 320-341.

Maslow, A. H. (1954). *Motivation and personality*. New York: Harper & Brothers.

Medina, T. R., Smith, S. N., & Long, J. S. (2009). Measurement models matter: Implicit assumptions and cross-national research. *International Journal of Public Opinion Research*, *21*(3), 333-361.

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

Meuleman, B., & Billiet, J. (2009). A Monte Carlo sample size study: How many countries are needed for accurate multilevel SEM? *Survey Research Methodology*, *3*(1), 45-58.

Meuleman, B., Davidov, E., & Billiet, J. (2009). Changing attitudes toward immigration in Europe, 2002-2007. A dynamic group conflict theory approach. *Social Science Research*, *38*(2), 352-365.

Muthén, B. O. (1985). A method for studying the homogeneity of test items with respect to other relevant variables. *Journal of Educational Statistics*, *10*(2), 121-132.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*, 557-585.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement*, *28*, 338-354.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, *22*(3), 376-398.

Muthén, L. K., & Muthén, B. O. (1998-2010). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.

Oort, F. J. (1992). Using restricted factor analysis to detect item bias. *Methodika*, *6*, 150-166.

Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling*, *5*, 107-124.

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

Saris, W. E., Satorra, A., & Sörbom, D. (1987). The detection and correction of specification errors in structural equation models. *Sociological Methodology*, *17*, 105-129.

Saris, W. E., Satorra, A., & van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling*, *16*, 561-582.

Schlüter, E., & Meuleman, B. (2009, June-July). *Measurement inequivalence as a source of useful information: Using multi-level SEM to explain why measurements are inequivalent*. Paper presented at the 3rd ESRA Conference, Warsaw, Poland.

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in Experimental Social Psychology*, *25*, 1-65.

Schwartz, S. H. (1994). Are there universal aspects in the content and structure of values? *Journal of Social Issues*, *50*, 19-45.

Schwartz, S. H., Melech, G., Lehmann, A., Burgess, S., Harris, M., & Owens, V. (2001). Extending the cross-cultural validity of the theory of basic human values with a different method of measurement. *Journal of Cross Cultural Psychology*, *32*, 519-542.

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. H. Poortinga (Eds.), *Individuals and cultures in multi-level analysis* (pp. 93-119). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Sörbom, D. (1978). An alternative to the methodology for analysis of covariance. *Psychometrika*, *43*, 381-396.

Steenkamp, J.- B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*, 78-90.

Stegmueller, D. (2011, July). *How many countries do you need to do multilevel modeling? A Monte Carlo experiment comparing frequentist and Bayesian approaches*. Paper presented at the 4th conference of the European Survey Research Association, Lausanne, Switzerland.

Steinmetz, H., Schmidt, P., Tina-Booh, A., Wieczorek, S., & Schwartz, S. H. (2009). Testing measurement invariance using multigroup CFA: Differences between educational groups in human values measurement. *Quality & Quantity*, *43*, 599-616.

United Nations Development Program. (2006). *Human Development Report 2006. Beyond scarcity: Power, poverty and the global water crisis*. Retrieved February 8, 2009, from http://hdr.undp.org/en/reports/global/hdr2006/chapters/

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*, 4-69.

Van der Veld, W., & Saris, W. E. (2011). Causes of generalized social trust: An innovative cross-national evaluation. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-cultural analysis: Methods and applications* (pp. 207-247). New York: Routledge.

Welkenhuysen-Gybels, J. (2003). *The detection of differential item functioning in Likert score items* (unpublished doctoral dissertation). Leuven, Belgium: KU Leuven.

Welkenhuysen-Gybels, J., van de Vijver, F., & Cambré, B. (2007). A comparison of methods for the evaluation of construct equivalence in a multi-group setting. In G. Loosveldt, M. Swyngedouw, & B. Cambré (Eds.), *Measuring meaningful data in social research* (pp. 357-372). Leuven, Belgium: Acco.