How to obtain comparable measures for cross-national comparisons

Cieciuch, Jan ; Davidov, Eldad ; Schmidt, Peter ; Algesheimer, René

Abstract: Comparisons of means or associations between theoretical constructs of interest in cross-national comparative research assume measurement invariance, that is, that the same constructs are measured in the same way across the various nations under study. While it is intuitive, this assumption needs to be statistically tested. An increasing number of sociological and social psychological studies have been published in the last decade in which the cross-national comparability of various scales such as human values, national identity, attitudes toward democracy, or religiosity, to name but a few, were tested. Many of these studies did not manage to fully achieve measurement invariance. In this study we review, in a nontechnical manner, the methodological literature on measurement invariance testing. We explain what it is, how to test for it, and what to do when measurement invariance across countries is not given in the data. Several approaches have been recently proposed in the literature on how to deal with measurement noninvariance. We illustrate one of these approaches with a large dataset of seven rounds from the European Social Survey (2002–2015) by estimating the most trustworthy means of human values, even when strict measurement invariance is not given in the data. We conclude with a summary and some critical remarks.

DOI: https://doi.org/10.1007/s11577-019-00598-7

Other titles: Wie kann man invariante Messungen in international vergleichender Forschung erhalten?

How to obtain comparable measures for cross-national comparisons

Jan Cieciuch*, University of Zurich and Cardinal Wyszynski University in Warsaw

jancieciuch@gmail.com

Eldad Davidov, University of Cologne and University of Zurich

e.davidov@uni-koeln.de

Peter Schmidt, University of Giessen

peter.schmidt@sowi.uni-giessen.de

René Algesheimer, University of Zurich

rene.algesheimer@business.uzh.ch

* Corresponding author

Number of words: 11,597

Number of characters (including references, tables, figures, and abstract): 77,743

Acknowledgments: The work of the first, second and fourth authors was supported by the University Research Priority Program Social Networks of the University of Zurich. The work of the third author was supported by the Alexander von Humboldt Polish Honorary Research Fellowship granted by the Foundation for Polish Science for the international cooperation between Peter Schmidt and Jan Cieciuch. The authors would like to thank Lisa Trierweiler for the English proof of the manuscript.

Bios

Cieciuch, Jan, 1974, Dr., University of Zurich and Cardinal Stefan Wyszynski University in Warsaw. Areas of research: structure and development of personality with focus on values, traits, personal identity, and well-being, also from a cross-cultural perspective.

Davidov, Eldad, 1971, Prof. Dr., University of Cologne and University of Zurich. President of the European Survey Research Association (ESRA) between 2015 and 2017. Areas of research: structural equation modeling especially applied to cross-cultural and longitudinal survey data. In his research he analyzes human values and attitudes toward immigrants or other minorities.

Schmidt, Peter, 1942, Prof. em. Dr., University of Giessen. Areas of research: foundations and applications of generalized latent variable models, and especially structural equation models. Applications include cross-country, repeated cross-sections and panel data. The substantive topics deal with values, the reasoned action approach, attitudes toward minorities, national identity and innovation. Together with A. Heath, E. Green, E. Davidov and A. Ramos, he was a member of the Question Design Team for the immigration module of the ESS 2014.

Algesheimer, René, 1973, Prof. Dr., University of Zurich. Director of the University Research Priority Program on Social Networks. Areas of research: how social structures shape individuals' behavior, and correspondingly how individuals' behavior reproduces social structures. More recently he has worked on the effects of social networks on influence processes, opinion diffusion, risk preferences, social learning, human values, and cooperation.

How to obtain comparable measures for cross-national comparisons

Abstract

Comparisons of means or associations between theoretical constructs of interest in cross-national comparative research assume measurement invariance, that is that the same constructs are measured in the same way across the various nations under study. While it is intuitive, this assumption needs to be statistically tested. An increasing number of sociological and social psychological studies have been published in the last decade in which the cross-national comparability of various scales such as human values, national identity, attitudes toward democracy, or religiosity, to name but a few, were tested. Many of these studies did not manage to fully achieve measurement invariance. In this study we review, in a nontechnical manner, the methodological literature on measurement invariance testing. We explain what it is, how to test for it, and what to do when measurement invariance across countries is not given in the data. Several approaches have been recently proposed in the literature on how to deal with measurement noninvariance. We illustrate one of these approaches with a large dataset of seven rounds from the European Social Survey (2002-2015) by estimating the most trustworthy means of human values, even when strict measurement invariance is not given in the data. We conclude with a summary and some critical remarks.

Keywords:

Exact and approximate measurement invariance; Alignment; Human values; European Social Survey

1. Introduction

Comparisons in cross-national comparative research (CNCR) assume that the same constructs are measured in the same way across the various nations under study. It is a very basic and intuitive assumption. However, this assumption becomes quite problematic in applied research because fulfilling this assumption is often not easy. Indeed, the methodological literature suggests that even the strict application of the same procedures of data collection and utilizing excellent translations of measurement instruments may not guarantee that the measurements are comparable and that cross-country comparisons based on these measurements would lead to meaningful results. Thus, while the paper of Goerres et al. in this section discusses research designs and case selection for CNCR, Meuleman presents analytical techniques to analyze such data, and Schmidt-Catran reviews techniques to deal with small and nonrandom country samples as well as unobserved heterogeneity in CNCR, our study focuses on how to make sure that our variables analyzed across the countries in CNCR are comparable. Three main questions arise: (1) What does it mean to measure the same construct in groups that differ in terms of language, history, culture, etc., or to measure the same construct over time?; (2) How can we be sure that we are measuring the same construct?; and (3) How can one obtain comparable scores across different groups under study? Answers to these questions are usually given in the framework of factor analysis and measurement models, where question items are used as reflective indicators for measuring latent variables that represent the theoretical constructs of interest that we want to compare. Before any cross-national comparisons are conducted, it is necessary to ensure that the same latent variables are measured in different countries, that respondents understand the items in a similar manner, and that they use the response scales in the same way. Meeting these three conditions allows us to speak of measurement invariance (Davidov et al. 2014).

We will present the topic of measurement invariance below, before going on to explain why it is important, what a test of measurement invariance requires and the logic behind it, how it can be tested across countries, and how and under which conditions its requirements may be released. We will demonstrate its procedure by examining the measurement invariance properties across countries and time points of human values measurements in the European Social Survey (ESS) between 2002 and 2015. We will then close with a summary and some concluding and critical remarks on the significance and indispensability of measurement invariance testing in cross-national research.

2. Measurement invariance: What is it, how do we test for it, and how do we deal with noninvariance?

2.1. Measurement invariance

Measurement invariance (or measurement equivalence) implies that using the same questionnaire in different groups (such as countries or at various points in time, or under different conditions) does measure the same construct in the same way (Chen 2008; Davidov et al. 2014; Horn and McArdle 1992; Millsap 2011). When measurement invariance is not established, comparisons between groups may not be meaningful because there is then no way to correctly determine whether observed differences across the groups are "true" or are only a methodological artifact (Chen 2008; Davidov et al. 2014). If a measurement is noninvariant, then findings on similarities or differences across groups may be misleading. It could well be the case that differences that are found between groups do not correspond to real differences, or that observed similarities would not reflect real similarities. For example, social desirability response bias may be stronger in one country than in another, a question may be unclear to many respondents in a certain country due to the cultural specificity of a question, or a construct might have a different meaning across different nations. Factors such as these might

lead to a different understanding of survey questions or to a different use of the response scale, thus rendering responses noncomparable across nations (Davidov et al. 2014).

2.2. How to test for measurement invariance

Measurement invariance can be tested empirically. The most commonly used method to test it is multigroup confirmatory factor analysis (MGCFA: Jöreskog 1971; see also Cieciuch and Davidov 2012, 2015). The test requires using latent variables with multiple measures (indicators). MGCFA assesses whether (1) the same measurement model is used in all groups, (2) factor loadings are the same across groups, (3) measurement intercepts are the same across the groups to be compared, and (4) residual variances are fixed to be equal across groups. The first condition is known as configural invariance, and satisfying it still precludes comparisons. The second condition is referred to as metric invariance. It implies that the scale intervals are the same across groups because the loadings are the same in each group. Satisfying the second condition allows comparing unstandardized regression coefficients and/or covariances across groups. The third condition is known as scalar invariance. Meeting it allows also comparing the latent means across groups meaningfully. Scalar invariance is the most restrictive model, since it requires both factor loadings and intercepts to be the same in all groups. However, when it is fulfilled, it implies that the researcher may carry on any comparison across the groups in question with confidence. The fourth condition is dubbed full uniqueness measurement invariance. This basically means that the explained variance for every item is the same across groups, in other words that the latent construct is measured identically across groups. When the error variances are not equal and complete uniqueness is not given, it implies that the items are measured with different amounts of error in different groups, but that one can still compare unstandardized regression coefficients and latent means across groups. Therefore, typical multigroup factor analysis generally only applies the first three steps.

Technically speaking, the test of measurement invariance in this framework involves setting cross-group constraints on parameters (loadings or loadings and intercepts) and comparing hierarchically more constrained models with less constrained ones (Davidov et al. 2014; Vandenberg and Lance 2000). These models are hierarchical in the sense that, at the configural level, all loadings and intercepts are freely estimated, while at the metric level loadings are constrained to be equal across groups, and at the scalar level both loadings and intercepts are constrained to be equal across groups. If the fit of the more highly constrained model does not deteriorate considerably, then one can assume that it is supported by the data and that the corresponding measurement invariance level is established.[1]

To illustrate, imagine respondents answering a survey question in two countries, Austria (A) and Belgium (B). Figure 1 describes three scenarios of associations between the latent construct of interest, on the X axis, and the response to a survey question measuring this construct of interest, on the Y axis. Each scenario describes the relation between the latent variable (X) and the question item (Y). Each line describes the association between the latent variable and the item in one country. The first scenario (a) illustrates configural invariance (and metric and scalar *non*invariance). In this scenario, both the factor loadings and the intercepts are different across countries, as evidenced by the different slopes and intersections with the Y axis. The second scenario (b) describes metric invariance (and scalar *non*invariance). In this scenario, the slopes are identical, and thus reflect the fact that factor loadings are the same in the two countries. However, the intercepts are not. The third

[1] Researchers examine global fit measures and perform chi-square difference tests to determine whether a more highly restricted model is supported by the data, that is if a higher level of invariance is given. However, based on a Monte Carlo study, Chen (2007) proposed an alternative to the chi-square difference test, which leads too easily to a rejection of measurement invariance. He proposed that metric noninvariance is indicated by a change smaller than 0.01 in the comparative fit index (CFI), supplemented by a change smaller than 0.015 in the root mean square error of approximation (RMSEA), or a change smaller than 0.03 in the standardized root mean square residual (SRMR) compared with the configural invariance model. To guarantee scalar invariance, Chen (2007) proposed to inspect whether the change in CFI is smaller than 0.01, the change in RMSEA is smaller than 0.015, or the change in SRMR is smaller than 0.01, when moving from a metric to a scalar invariance model for sample sizes larger than 300 per group.

scenario (c) presents metric and scalar invariance. In this case, both the loadings and the intercepts are equal across countries (which is why one can observe only a single line). The figure makes it clear that, in the two scenarios (a) and (b), observed scores are different, even when the true score (on the X axis) is the same in both countries. Only in scenario (c) is it possible to observe the same score when the true score is identical in both countries.

Please insert Figure 1 here

2.3. Dealing with noninvariance

While some studies were successful in establishing high levels of measurement invariance across groups (e.g. Davidov and Siegers 2010), applied research in the last decade has unfortunately shown that many scales fail to display comparability across countries or time. For example, Aleman and Woods (2016) and Sokolov (2018) recently showed that the Inglehart value scales are not comparable across all countries (but see Welzel and Ingelhart 2016); Ariely and Davidov (2010) found that public support for democracy cannot be compared across countries in the World Value Survey (WVS); Lomazzi (2018) demonstrated that gender-role attitudes are not comparable across all WVS countries; Davidov et al. (2008) showed that means of human values in the ESS may not be compared across all countries; Rudnev et al. (2018) found that the means of Seeman's alienation scale are not comparable cross-nationally; Davidov (2009) discovered that the scale means of nationalism and patriotism are not comparable across countries participating in the International Social Survey Program (ISSP); and Coromina and Davidov (2013) concluded that social and political trust are not always comparable across countries and/or time points (see also Marsh et al. 2017). To tackle this problem, several solutions were proposed including technical procedures and theoretical approaches. What most of these approaches have in common is that they suggest releasing some of the strict constraints required by scalar invariance models while not

compromising on the scales' comparability. Figure 2 presents an overview of the main recommendations in the form of a decision tree. We will discuss each of these approaches below.

Please insert Figure 2 here

Figure 2 depicts several analytical procedures that can be applied when noninvariance is obtained. First, several researchers suggested that instead of finding a model that meets measurement invariance requirements, noninvariance may be used as a unique opportunity and a useful source of information about cross-country differences (assuming that the theoretical model and measurement instrument are sound). They suggested that multilevel structural equation modeling (MLSEM) may be employed to explain why measurement invariance is not given (Davidov et al. 2012, 2016; Jak et al. 2013). MLSEM differentiates between the measurement model on the individual level and the measurement model on the country level. Researchers can use contextual variables such as the economic conditions in a country, policies, or the Human Development Index to try to explain, in a theoretically driven way, the lack of invariance of specific measurement items.

When the goal is, however, to identify a measurement model where measurement invariance holds to be able to conduct a meaningful cross-country comparison, then finding the reasons for noninvariance might not be sufficient. MLSEM may explain noninvariance, but does not necessarily ratify or correct for it. Thus, other researchers have proposed to release the strict constraints of measurement invariance testing so that the measurement invariance model fits the data better. They suggested that less strict models would be more realistic but still good enough for conducting meaningful comparisons. Some of them suggested testing for partial (metric or scalar) measurement invariance, rather than for full

(metric or scalar) invariance. These researchers argued that partial invariance may be sufficient for meaningful comparisons (Byrne et al. 1989; Steenkamp and Baumgartner 1998). Partial invariance is established when the parameters of at least two indicators (loadings at the metric level and loadings plus intercepts at the scalar level of the measurement) are equal across groups. In other words, the researcher identifies those items which are very different across groups, and releases them while ensuring that at least two items per scale have equal loadings and intercepts. Whereas this approach has been applied quite frequently by substantive researchers, it has also been criticized as insufficient to guarantee meaningful comparisons (see e.g. De Beuckelaer and Swinnen 2018; Steinmetz 2018). As a result, methodologists have recently come up with newer proposals on how to deal with the lack of measurement invariance.

Testing for full or partial measurement invariance, which we described above, is considered an *exact* approach, because testing for either full or partial measurement invariance assumes that at least some of the parameters (loadings and/or intercepts) are *exactly* equal across groups. Recently, Muthén and Asparouhov (2013) suggested replacing the requirement of exact equality of measurement parameters with the requirement of an *approximate* equality of measurement parameters. They argued that approximate invariance may be sufficient for meaningful country comparisons (Muthén and Asparouhov 2013; Van de Schoot et al. 2013). In other words, they suggested that it is sufficient when the parameters (factor loadings or intercepts) are more or less (rather than exactly) equal.

Tests for approximate measurement invariance can be performed in the Bayesian framework, where loadings and intercepts are treated as variables with a specific distribution. The parameters of this distribution (means and variance) are known as *priors*, and can be defined by the researcher based on previous knowledge or assumptions (Muthén and Asparouhov 2013). In the exact measurement invariance approach, loadings and intercepts are

constrained to be exactly equal, and consequently the *differences* between them are assumed to be zero by definition, with a zero variance. In the approximate approach, one assumes that the cross-country mean of the *differences* between loadings or intercepts equals zero, but that small variations for these mean differences are allowed. The amount of the variation that is allowed is indicated by the variance of the cross-country parameter differences that can be imposed on the model. Several simulation studies have shown that variances equal to 0.01 or 0.05 in the distribution of the cross-country differences in loadings or intercepts probably do not bias substantive conclusions for comparative research (Muthén and Asparouhov 2013; Van de Schoot et al. 2013). More complete research on the tolerated level of variability of the differences between parameters is still missing, so it is still not clear in the literature which parameter differences may be tolerated and which may be too large to guarantee meaningful comparisons.

The test for approximate measurement invariance (implemented in the Mplus structural equation modeling software package: Muthén and Muthén 1998-2014, or in lavaan: Merkle and Rosseel 2016), provides researchers with two types of measures with which to assess the quality of the models. The first type measures the global fit of the model and includes the posterior predictive p-value (ppp) and the credibility interval (CI) for the difference between the observed and replicated chi-square values (but see Marsh et al. 2017 for the newly developed pppp measure, which is yet to be implemented in a commercial software package). According to Muthén and Asparouhov (2013) and Van de Schoot et al. (2013), the Bayesian model fits the data when the ppp is not significant (larger than zero) and the CI contains a zero.

The second type of measure obtained in approximate measurement invariance testing is more detailed. It includes a list of all parameters that are noninvariant in each group and the significance of the deviation. Using this list, researchers can identify items that are

particularly noninvariant, that is, deviate from invariance in many groups. In the next step, researchers may decide whether to drop these items, drop countries that have a particularly high number of noninvariant items, or increase the tolerated variance of parameter differences in the model (see e.g. Davidov et al. 2015). When following this strategy it is indeed, as indicated above, not clear to what extent this variance may be increased under different conditions without placing the meaningfulness of cross-country comparisons at risk.

Another approximate approach which has recently been proposed in the literature is alignment optimization (Asparouhov and Muthén 2014). This approach allows, under certain conditions, an unbiased comparison of means using an optimization process, even in the presence of noninvariance. The optimization process computes the most trustworthy latent means (Asparouhov and Muthén 2014; Cieciuch et al. 2018; Muthén and Asparouhov 2014, 2017) *without* constraining loadings and intercepts to be equal across groups (i.e. by using a configural invariance model only) (Asparouhov and Muthén 2014; Muthén and Asparouhov 2014). Thus, the means are estimated while taking into account real differences in loadings and intercepts. The alignment optimization method discovers the most optimal measurement invariance pattern, in which a relatively small number of large noninvariant parameters – and many approximately invariant parameters – are present, rather than imposing exact equality constraints on all parameters. Asparouhov and Muthén (2014) compare this procedure to a rotation in exploratory factor analysis which simplifies the loading matrix without modifying the model fit. The alignment optimization procedure identifies the noninvariant parameters (loadings and intercepts). It can be performed on any multiple group model, and also when measurement properties are significantly different across groups.

Muthén and Asparouhov (2014) proposed a rule of thumb for determining when it is safe to continue performing the mean comparisons. They suggested that such a comparison is meaningful when up to about 25% of the parameters (factor loadings and intercepts) are

noninvariant. However, this recommendation is based on a very limited set of simulation studies. It is still unclear whether this rule of thumb is too strict or too liberal. It is also not clear whether 25% of all measurement parameters (factor loadings and intercepts) may be noninvariant, or whether this rule of thumb should be applied separately for each set of parameters. Further simulation studies are required in order to determine how many parameters may be noninvariant without risking the meaningfulness of the mean comparisons across countries when using alignment. Notwithstanding these limitations, the alignment approach is particularly useful for substantive research because it is easier to apply than other tests of measurement invariance, especially when the number of countries to be compared is large.

In the next section we will illustrate the procedure on the human values measurements in the ESS. We will perform a large-scale test of the measurement invariance properties of human values across all countries which participated in seven rounds of the ESS. This is, to the best of our knowledge, the largest measurement invariance test applied to survey data using the alignment procedure (for a similarly large study using Alignment, see Munck et al. 2017). Human values scores obtained using this procedure may be potentially relevant when researchers are interested in comparing value scores across ESS countries *and* time points.

3. Measurement (non)invariance of human values as measured in the ESS

Value preferences are considered in many sociological and social-psychological studies to be a dimension of major importance to describe persons, groups, and societies, and to explain attitudes, intentions, and behavior (Durkheim 1897/1964; Hitlin and Piliavin 2004; Hofstede 2000; Inglehart and Baker 2000; Kluckhohn 1951; Rokeach 1973; Schwartz 1992; Weber 1905/1958). Although several researchers developed different value theories and proposed various scales to measure them, the circular model of values proposed by Schwartz (1992,

Schwartz et al. 2012; Steinmetz et al. 2012) is probably the one most frequently applied in the social sciences. Measures of the Schwartz human values scale are included in all seven rounds of the ESS (Schwartz 2003)[2]. The model has received empirical support in many studies worldwide (Bilsky et al. 2011; Steinmetz et al. 2012; Rudnev et al. 2018), in samples of adults and children (Cieciuch et al. 2016; Döring et al. 2015) using both cross-sectional and longitudinal data. There is no doubt that including value measurements in the ESS offers a unique opportunity to analyze value priorities and change in many European countries. However, a methodological precondition for performing such analyses meaningfully is to ensure that the value measurements in the ESS are invariant across countries and points in time, and therefore comparable. Obtaining the most trustworthy value means in the ESS is thus of paramount importance for substantive research. Below, we first present the value model to be examined and briefly review previous results on measurement invariance testing of the ESS value scale. Next, we build on previous results and apply the alignment optimization to the values in order to obtain the most trustworthy value means in the ESS.

3.1. Basic human values – the circular model and the value measurement in the ESS

In the circular model proposed by Schwartz (1992; Schwartz et al. 2012), values are defined as broad, transsituational goals that vary in importance and serve as guiding principles in the life of a person or group. The number of basic values according to the theory is limited. People or groups differ in their *hierarchy* of values rather than in the set of values they consider important. The main claim of Schwartz' (1992; Schwartz et al. 2012) circular model is that all values can be located on the circle according to the motivation they express. Neighboring values are based on a similar motivation, and can be pursued in the same action.

[2] Measures of the Schwartz values are also included in other international surveys such as the World Values Survey or the U.S. General Social Survey (for further details, see http://www.worldvaluessurvey.org/wvs.jsp; http://www.norc.org/Research/Projects/Pages/general-social-survey.aspx).

Values located on opposite sides of the circle express conflicting motivations in the sense that they cannot be pursued concurrently when performing the same behavior.

The value circle represents a circular motivational continuum. Thus, the circle can be divided in many ways according to the research goals and the measurement instrument used. Traditionally, the value circle was divided into ten values that form four higher-order values, as presented in Figure 3 (the first two internal circles) (Schwartz 1992). The four higher-order values describe self-transcendence vs. self-enhancement values and conservation vs. openness to change values. The refined model (Schwartz et al. 2012) enables a division into 19 more narrowly defined values. At the same time, the value circle can also be divided into more general and broadly defined values, which are represented by the third and the fourth circles in Figure 3. One division differentiates between anxiety-free growth values which oppose anxiety avoidance self-protection values. Another division differentiates between sociallyfocused values vs. values with a personal focus.

Please insert Figure 3 here

Schwartz developed several methods to measure the ten values and the four higher-order values. The Portrait Value Questionnaire (PVQ) is the scale that is used most frequently (for a review, see Schwartz and Cieciuch 2016). The basic version of the questionnaire consists of 40 items (Schwartz et al. 2001) that describe other people. Respondents have to assess the similarity between themselves and the people described. A shortened version of the questionnaire consisting of only 21 items (PVQ-21: Schwartz 2003) was included in the ESS from the very beginning. Each item in the PVQ-21 is composed of two sentences which describe a portrait from a male or female perspective. The portraits contain goals, aspirations or desires that point to the importance of a value. For each item, the respondents in the ESS

answer the question “How much like you is this person?”, with a response scale ranging from 1 (not like me at all) to 6 (very much like me). The items included in the ESS to measure the ten values and the four higher-order dimensions are presented in Table 1.

Please insert Table 1 here

3.2. Measurement invariance of values in the ESS and dealing with noninvariance: A review and an illustration

Values were measured in the ESS using the same questionnaire translated into different languages and by applying mostly face-to-face interviews (for exceptions and documentation of the data collection procedures, see the ESS website at www.europeansocialsurvey.org). This careful procedure is however not sufficient to ensure the comparability of values in the ESS, and their measurement invariance properties have to be examined before using their ESS measures in comparative research meaningfully. Below we will follow the steps depicted in Figure 2 that guide the assessment of measurement invariance and how to deal with noninvariance.

3.2.1 Lack of exact scalar measurement invariance

First, in 2008, Davidov et al. published a seminal paper testing for measurement invariance of values across 20 countries using data from the 1st round of the ESS. The findings were also replicated for the 2nd and 3rd rounds of the ESS (Davidov 2008, 2010). The two main results obtained by Davidov and colleagues are as follows: (1) Only seven values can be differentiated in most of the countries. Specifically, three pairs of adjacent values need to be unified: power with achievement, benevolence with universalism, and conformity with tradition. (2) Only metric invariance is established for the seven values. Scalar invariance across countries was not supported by the data. Thus, the findings suggested that cross-

country mean comparisons of the values as measured in the ESS may be problematic. However, Davidov (2008, 2010) demonstrated that values displayed scalar invariance over time, suggesting that their means may be used for longitudinal comparisons in the countries of interest. Figure 2 presented various procedures to deal with measurement noninvariance. In the next steps, these procedures were applied in value research in a number of different ways that we briefly describe below.

3.2.2 Refining the theory

In 2012, Schwartz and colleagues used information about noninvariance to improve the model and its measurement instrument. They refined the value theory to include 19 instead of ten values, and developed a new 57-item scale that was better suited to measuring the single values (Knoppen and Saris 2009; Beierlein et al. 2012). This refined theory did not contradict its older version. After all, the original theory suggested that one may divide the value circle into more or less specific values depending on the measurement instruments one has (Cieciuch et al. 2013). Cieciuch et al. (2014b) and Cieciuch et al. (2014a) demonstrated that this version of the theory and its measurement instrument can distinguish between all single values, and the refined instrument possesses much better cross-country measurement invariance properties than the ESS scale does.

3.2.3 Using noninvariance as a source of information on cross-cultural measurement differences

Another approach in Figure 2 proposed using findings of cross-country noninvariance of the ESS value scale as an important source of information on country differences. Applying multilevel structural equation modeling, Davidov and colleagues (2012) showed how to explain the noninvariance of a specific item measuring universalism which tapped into the importance of protecting the environment. This item was particularly noninvariant across

European countries. The authors found that it was endorsed more strongly by residents of European countries that had a lower Human Development Index score. These populations apparently considered clean water and air to be a matter of survival and health, whereas these aspects appear to be taken for granted by individuals residing in more developed countries.

3.2.4 Applying approximate measurement invariance testing techniques

Finally, the third approach as presented in Figure 2 questioned the procedure of testing for exact measurement invariance of human values. Referring to new developments in Bayesian analysis, it raised the possibility that previous measurement invariance tests on the ESS value scale may have been too strict. After all, as discussed earlier, a small degree of noninvariance may not necessarily bias substantive conclusions in comparative studies. Thus, instead of relying on exact measurement invariance, Cieciuch et al. (2017) and Zercher et al. (2015) proposed that the ESS value scale be subjected to tests of approximate measurement invariance. They applied approximate measurement invariance procedures based on Muthén and Asparouhov (2013) and Van de Schoot et al. (2013), who suggested that factor loadings and intercepts need not necessarily be *exactly* equal across countries. Instead, they may be almost equal.

Zercher and colleagues (2015) applied, for the first time, the approximate measurement invariance approach to values measured in the ESS data. They focused on only one value – universalism –, and ran the approximate measurement invariance test across 15 countries which participated in all six available ESS rounds simultaneously, resulting in a comparison of (15 x 6 =) 90 groups. They showed that, whereas scalar measurement invariance in the (traditional) exact approach was established across 37 groups, approximate measurement invariance could be established across no fewer than 73 groups, thus challenging previous findings on lack of invariance. Hence, whereas the universalism value

was not measurement invariant across all country/time-point combinations, it was comparable across most of them.

Cieciuch and colleagues (2017) went one step further and subjected the full ESS value scale to an approximate invariance test for all values. However, due to the highly complex nature of the model, they did not run a simultaneous test across all 15 countries and six measurement time points of the ESS. Instead, they tested for approximate measurement invariance across the 15 countries *within* each ESS round. Furthermore, given that it was not possible to measure ten values, Cieciuch et al. conducted the approximate measurement invariance test *separately* on each higher-order value using the magnifying glass strategy (Cieciuch and Schwartz 2012) as illustrated in Figure 4. This approach helped to avoid introducing cross-loadings which are inherent in the theory when all values are used simultaneously in a single model. Cieciuch et al. (2017) established approximate measurement invariance successfully for the higher-order values in most countries.[3] While these tests resulted in higher levels of measurement invariance than previous tests did, it is yet to be examined whether the tests were too liberal in the sense that they allowed too much variability of measurement parameters across countries. Indeed, researchers still need to determine how much variability may be allowed in approximate measurement invariance testing.

Please insert Figure 4 here

3.2.5 Applying the alignment optimization

We wish to illustrate below another technique presented in Figure 2 (building on Cieciuch et al. 2017), namely the alignment optimization. This is also a method which allows for

[3] The authors excluded the value 'hedonism' from this model. According to the theory, this value is located between openness to change and self-enhancement. Including this value in either of the models resulted in a significant reduction in model fit.

approximate rather than exact measurement invariance. However, instead of imposing exact or approximate equality constraints on the measurement parameters, it searches for the pattern of loadings and intercepts that minimize noninvariance. As a result, it produces the most trustworthy value means from the data[4]. For the illustration we used the same human values models as those presented in Figure 4, and analyzed 15 countries which participated in the first seven ESS rounds (2002/2003, 2004/2005, 2006/2007, 2008/2009, 2010/2011, 2012/2013, and 2014/2015). We utilized the Mplus software package Version 7.3 (Muthén and Muthén 1998-2014). The syntax of the analysis is available from the first author upon request.

Table 2 presents the number of respondents in each round and country included in the analysis. We followed the recommendations provided on the ESS website, and only considered respondents with no more than five missing values and no more than 16 identical responses for the 21 value items. As a result, the analysis included a total of 175,452 respondents. The remaining item nonresponse was dealt with by using the full information maximum likelihood (FIML) procedure (see Schafer and Graham 2002). The ESS website (www.europeansocialsurvey.org) provides documentation about the data collection procedure and permits the data and accompanying material to be downloaded. Table 3 presents the global fit measures for the configural invariance model of each higher-order value across 105 groups (15 countries x 7 rounds). The global fit measures suggested that the four models had a good fit to the data. All 1,995 factor loadings of the higher-order values as depicted in

[4] The scale of latent variables is unknown, and hence their variance is also unknown (by definition, these variables are unobserved). Therefore, in order to identify the model, researchers need to apply some restriction for the estimation: either restricting the variance of the latent variable to an arbitrary value (typically it is then restricted to 1 in all groups), or fixing the scale of the latent variable by restricting the factor loading of one of the items (the so-called anchor item) to 1 in all groups. When doing so, it is important to guarantee that such a restriction fits the data at hand. In the former case, the restriction implies an implicit assumption that the latent variance is equal across groups. In the latter case, the restriction implies that the factor loading of the anchor item is indeed equal across groups. In both cases, researchers need to make sure that the assumption holds, for example by inspecting which of these parameters (factor loading of one of the items or the latent variable variance) are indeed most similar across groups, and choose the restriction which best corresponds with the data at hand (see also Brown 2015, p.271).

Figure 3 (six for conservation, five for self-transcendence, and four for openness to change and self-enhancement, respectively, in 105 country/time combinations) were significant, and almost all of them were higher than 0.4 (Brown 2015). Only very few loadings (i.e. ten, which corresponded to about 0.5% of the total number of loadings) were slightly lower than 0.3. We nevertheless retained these items because these cases were very few in number, and we wanted our measurement models to correspond to the theory. Further information on the factor loadings and the measurement models may be obtained from the first author upon request.

Please insert Table 2 here

Please insert Table 3 here

Table 4 presents the number of noninvariant loadings and intercepts reported in the Mplus output (Muthén and Muthén 1998-2014). This information is important to determine whether the means computed by the alignment procedure are trustworthy. As previously indicated, Muthén and Asparouhov (2014) suggested that a cutoff criterion of about 25% of noninvariant parameters (factor loadings and intercepts) may not be exceeded. The average amount of noninvariance was lowest for self-transcendence (28%), followed by conservation (29%) and self-enhancement (33%), and was highest for openness to change (34%), and therefore slightly above the 25% cutoff criteria suggested by Muthén and Asparouhov (2014). As displayed in Table 4, noninvariance was particularly evident for intercepts. About half of the intercepts for the four higher-order values were not invariant according to the output. This finding corresponds to previous findings testing for measurement invariance of the ESS

values in which factor loadings were rather invariant but intercepts were not.[5] Therefore, although alignment provides the most trustworthy means possible with the data at hand, the findings for the means, and particularly those for openness to change and self-enhancement, where the number of noninvariant intercepts exceeded 50%, should be treated with great caution.

Please insert Table 4 here

Tables A1-A4 in the Appendix present the means for each higher-order value in each combination of countries/rounds. Latent means are meaningless per se, and should be interpreted in comparison to other country means. Countries with more than 25% noninvariant parameters (factor loadings and intercepts) are marked with an asterisk (*).[6] Furthermore, the tables present the country rankings within each round. This makes it possible to conclude from each table how, at some measurement time point, a country compares to any other country at the same or at any other time of measurement. For example, Table A1 shows that Poland displays the highest conservation value means in the first two rounds. In Round 7, Poland displays a higher mean compared to its measures in the first two rounds. However, it now ranks second (rather than first) in this last round because Slovenia is even more conservative in this round. Slovenia displays a large increase in its conservation scores between the 1st and 7th rounds, increasing from -0.877 to -0.089. As another example, turning to Table A2, in which self-transcendence scores are presented, we see that

[5] Many studies evidenced that intercepts were not equal across groups, and that it was easier to guarantee equal factor loadings than equal intercepts when comparing different countries (see e.g. Davidov et al. 2014). In other words, it was often easier to establish metric invariance than scalar invariance. Different intercepts may also reflect different country-specific survey strategies, which in turn may result in different response patterns across countries.

[6] Tables that display more highly specific information about the (non)invariance pattern for each higher-order value may be obtained from the first author on request. They present the number of noninvariant loadings and intercepts for each item and country. One way to estimate the amount of bias, discussed in Oberski (2014), is to perform a sensitivity analysis.

Switzerland considerably increased its self-transcendence scores when moving from the 1st round (-0.149) to the 7th round (0.187). However, even though Switzerland ranked highest in self-transcendence in the 1st round, it comes in the third place in the 7th round. The reason is that other countries displayed even more pronounced increases in self-transcendence during this period. Researchers who are interested in specific countries and time points may examine the scores for specific values in order to draw conclusions about value change and value development in countries. Thus, this illustration demonstrates that researchers can quite easily examine measurement invariance properties in complex and large cross-national datasets – such as in the large-scale investigation of human values considered here – and can estimate how the countries' mean scores compare with one another. In the final section, we will first reflect on the different approaches that can be implemented to test for measurement invariance discussed so far, and then consider the extent to which we may rely on mean scores when exact measurement invariance is not given by the data.

4. Summary and discussion

The last decade has witnessed an increase in the number of published studies that included the testing of the measurement invariance of various scales. This is an important development in the literature. After all, in cross-national comparative research, meaningful comparisons of means or associations between theoretical constructs that are of interest can only be performed when the same constructs are measured in the same way across the various nations under study. This assumption is also referred to as measurement invariance (or measurement equivalence). It must be satisfied in order to draw any meaningful conclusions in CNCR settings, either for direct comparisons of means or associations, or for multilevel analysis

which implicitly assumes that scores are comparable across the units of analysis (e.g. countries).[7]

There are various statistical procedures to perform measurement invariance tests, but the most common is the multiple group confirmatory factor analysis. In this study we reviewed, in a nontechnical manner, the methodological literature on measurement invariance testing. We explained what it is, how to test for it, and what to do when measurement invariance across countries is not given in the data.

Indeed, in many studies where measurement invariance was examined, it was not possible to achieve sufficient levels of invariance. Failing to reach measurement invariance may threaten the meaningful interpretation of comparisons across nations. Lack of measurement invariance could result in methodological artifacts that may be responsible for misleading differences observed between country scores. At the same time, similarities observed across nations when measurement invariance is not present might also be misleading, because they could mask true differences that cannot be observed due to a lack of comparability.

Several approaches on how to deal with measurement noninvariance, including testing for partial (rather than full) invariance, approximate (rather than exact) invariance and alignment, have been recently proposed in the literature. Alignment may be particularly interesting for applied researchers because it allows the most trustworthy means to be estimated in the data, even when measurement invariance is not present, since it provides researchers with the tools to assess whether and to what extent the scores are nevertheless comparable. It is also relatively easy to apply when the number of groups is very large.

[7] Methodologists also discuss the topic of isomorphism, which refers to equivalent construct meaning across levels of analysis. In other words, it refers to the presence or absence of measurement invariance across levels, for example across individuals and countries. However, examining isomorphism in cross-national data settings is beyond the scope of the present study (for a further discussion, see e.g. Guenole 2016; Muthén 1994; or Ruelens et al. 2016).

We illustrated the use of the alignment approach with a large dataset consisting of seven rounds from the ESS (2002-2015) by estimating the most trustworthy means of higher-order human values in Schwartz' (1992) model across all available ESS rounds, even when strict measurement invariance was not given in the data. We included in the illustration the 15 countries that participated in seven ESS rounds (i.e. 105 groups). Unfortunately, it was not possible to establish measurement invariance of the values across all ESS countries and points in time. The alignment procedure nevertheless allowed us to estimate the means – even in the absence of measurement invariance. The scores revealed a significant variability of the human values scores both across countries and over time. Some changes in country rank order were found across rounds for all values. We also observed that in many cases, countries ranking highest or lowest on specific values tended to remain in these positions across some of the rounds. However, there were several exceptions. The findings therefore suggest that while values are rather stable (in terms of country rankings), they did change slowly over time. Researchers interested in investigating specific values in specific countries and points in time may repeat the analysis for their specific countries of interest.

Our illustration underlines several limitations in measurement invariance testing in general, and in the alignment procedure in particular. While it is obvious that measurement invariance is a necessary condition for meaningful cross-national comparison, empirical studies frequently demonstrate that full invariance cannot be reached. As a result, based on simulation studies, recent developments have proposed relaxing some of the parameter equality requirements in measurement invariance testing by allowing for partial or approximate (rather than full or exact) invariance. These approaches suggest freeing some of the measurement parameters (factor loadings and/or intercepts) or only requiring the measurement parameters to be approximately (rather than exactly) equal. However, it is still to be studied to what extent such relaxations of the requirements in the statistical test are

legitimate. It is not yet clear how many equality constraints may be freed, and how many measurement parameters must remain equal across countries in the partial invariance test. It is also not yet fully clear how much variability in the differences between the measurement parameters may be allowed in the approximate invariance test. Finally, it is not yet known whether one may allow for more or less than 25% of the measurement parameters to differ in the alignment optimization procedure. After all, if excessively generous tolerances are applied in the tests, one may run the risk of rendering the scores nonequivalent. Indeed, whereas alignment optimization provides the most trustworthy means, even when several measurement parameters are not invariant, it still does not guarantee that all means are comparable. Thus, it is desirable to complement such more liberal tests with robustness tests in order to inquire whether freeing certain parameters leads to invalid conclusions (Oberski 2014). Notwithstanding these limitations, applied sociologists and social psychologists are encouraged to perform the analyses that we presented above. Researchers can be confident that their scores are comparable cross-nationally only after measurement invariance tests have been performed.

Measurement invariance is important not only for cross-country comparisons, but also for multilevel modeling. A multilevel analysis relies on the assumption that country-level effects (of, e.g., composite scores of trust or threat due to immigration) are comparable when estimating a random slope. It also assumes that means are comparable when one estimates a random intercept. Thus, at least partial metric invariance is necessary for estimating random slopes, and at least partial scalar invariance is needed for estimating random intercepts. Indeed, nearly all multilevel analyses that use composite scores of multiple indicators assume but do not test the assumption of measurement invariance of these scores. To the best of our knowledge, there is no simulation study which estimates whether and to what extent multilevel analysis findings are biased if measurement invariance across groups is absent.

Evaluating the implications of the lack of measurement invariance for estimating multilevel models using simulation studies is an important direction for future research.

Two final words of caution: First, one should take into account that even if metric and scalar invariance are given, there might be differences in meaning which cannot all be detected by quantitative techniques (Meitinger 2017). In such cases, one could consider applying mixed-methods approaches to explain instances of measurement noninvariance by combining measurement invariance tests with different qualitative techniques. Second, it might be important to check the heterogeneity of all the national population samples, as the intra-country mean differences might be greater than the inter-country differences (Magun et al. 2016; Schmidt-Catran in this volume). We hope that our review and illustration can provide researchers with tools to address the methodological challenges of comparability when using diverse scores in cross-national comparative research settings.

References

Aleman, Jose, and Dwayne Woods. 2016. Value orientations from the World Value Survey: How comparable are they cross-nationally? *Comparative Political Studies* 49:1039-1067. doi:10.1177/0010414015600458.

Ariely, Gal, and Eldad Davidov. 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.

Asparouhov, Tihomir, and Bengt O. Muthén. 2014. Multi-group factor analysis Alignment. *Structural Equation Modeling* 21:1-14. doi:10.1080/10705511.2014.919210.

Beierlein, Constanze, Eldad Davidov, Peter Schmidt, Shalom H. Schwartz, and Beatrice Rammstedt. 2012. Testing the discriminant validity of Schwartz' Portrait Value

Questionnaire items—A replication and extension of Knoppen and Saris (2009). *Survey Research Methods* 6:25–36.

Bilsky, Wolfgang, Michael Janik, and Shalom H. Schwartz. 2011. The structural organization of human values—Evidence from three rounds of the European Social Survey (ESS). *Journal of Cross-Cultural Psychology* 42:759-776. doi:10.1177/0022022110362757.

Brown, Timothy A. 2015. *Confirmatory factor analysis for applied research*. New York: Guilford Press.

Byrne, Barbara M., Richard J. Shavelson, and Bengt O. Muthén. 1989. Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin* 105:456-466. doi:10.1207/s15328007sem1102_8.

Chen, Fang F. 2007. Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling* 14:464-504. doi:10.1080/10705510701301834.

Chen, Fang F. 2008. What happens if we compare chopsticks with forks? The impact of making inappropriate comparison in cross-cultural research. *Journal of Personality and Social Psychology* 95:1005-1018. doi:10.1037/a0013193.

Cieciuch, Jan, Shalom H. Schwartz, and Michele Vecchione. 2013. Applying the refined values theory to past data: What can researchers gain? *Journal of Cross-Cultural Psychology* 44:1215-1234. doi:10.1177/0022022113487076.

Cieciuch, Jan, and Eldad Davidov. 2012. A comparison of the invariance properties of the PVQ-40 and the PVQ-21 to measure human values across German and Polish samples. *Survey Research Methods* 6:37-48.

Cieciuch, Jan, and Eldad Davidov. 2015. Establishing measurement invariance across online and offline samples. A tutorial with the software packages Amos and Mplus. *Studia Psychologica* 15: 83-99. doi:10.21697/sp.2015.14.2.06.

Cieciuch, Jan, Eldad Davidov, and René Algesheimer. 2016. The stability and change of value structure and priorities in childhood: A longitudinal study. *Social Development* 25:503-527. doi:10.1111/sode.12147.

Cieciuch, Jan, Eldad Davidov, René Algesheimer, and Peter Schmidt. 2017. Testing for approximate measurement invariance of human values in the European Social Survey. *Sociological Methods & Research* in press. doi:10.1177/0049124117701478.

Cieciuch, Jan, Eldad Davidov, and Peter Schmidt. 2018. Using alignment optimization in establishing measurement invariance. In *Cross-cultural analysis: Methods and applications, 2nd edition*, eds. Eldad Davidov, Peter Schmidt, Jaak Billiet, and Bart Meuleman. New York: Routledge Taylor & Francis Group.

Cieciuch, Jan, Eldad Davidov, Peter Schmidt, René Algesheimer, and Shalom H. Schwartz. 2014. Comparing results of an exact versus an approximate (Bayesian) measurement invariance test: A cross-country illustration with a new scale to measure 19 human values. *Frontiers in Psychology* 982:1-10. doi:10.3389/fpsyg.2014.00982.

Cieciuch, Jan, Eldad Davidov, Michele Vecchione, Constanze Beierlein, and Shalom H. Schwartz. 2014a. The cross-national invariance properties of a new scale to measure 19 basic human values. A test across eight countries. *Journal of Cross-Cultural Psychology* 45:764-779. doi:10.1177/0022022114527348.

Cieciuch, Jan, Eldad Davidov, Michele Vecchione, and Shalom H. Schwartz. 2014b. A hierarchical structure of basic human values in a third-order confirmatory factor analysis. *Swiss Journal of Psychology* 73(3):177-182. doi:10.1024/1421-0185/a000134.

Cieciuch, Jan, and Shalom H. Schwartz. 2012. The number of distinct basic values and their structure assessed by PVQ-40. *Journal of Personality Assessment* 94:321-328. doi:10.1080/00223891.2012.655817.

Cieciuch, Jan, Shalom H. Schwartz, and Eldad Davidov. 2015. Values, social psychology of. In *International Encyclopedia of the Social & Behavioral Sciences, 2nd edition, v. 25*, ed. James D. Wright, 41–46. Oxford: Elsevier.

Coromina, Lluis, and Eldad Davidov. 2013. Evaluating measurement invariance for social and political trust in Western Europe over four measurement time points (2002-2008). *ASK Research & Methods* 22(1):35-52.

Davidov, Eldad. 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* 2:33-46.

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

Davidov, Eldad. 2010. Testing for comparability of human values across countries and time with the Third Round of the European Social Survey. *International Journal of Comparative Sociology* 51:171-191. doi:10.1177/0020715210363534.

Davidov, Eldad, Jan Cieciuch, Peter Schmidt, Bart Meuleman, and René Algesheimer. 2015. The comparability of measurements of attitudes toward immigration in the European Social Survey: Exact versus approximate measurement equivalence. *Public Opinion Quarterly* 79: 244-266. doi:10.1093/poq/nfv008.

Davidov, Eldad, Hermann Dülmer, Elmar Schlueter, Peter Schmidt, and Bart Meuleman. 2012. Using a multilevel structural equation modeling approach to explain cross-cultural measurement noninvariance. *Journal of Cross-Cultural Psychology* 43:558-575. doi:10.1177/0022022112438397.

Davidov, Eldad, Hermann Dülmer, Jan Cieciuch, Anabel Kuntz, Daniel Seddig, and Peter Schmidt 2016. Explaining measurement nonequivalence using multilevel structural

equation modeling: The case of attitudes toward citizenship rights. *Sociological Methods & Research.* doi:10.1177/0049124116672678.

Davidov, Eldad, Bart Meuleman, Jan Cieciuch, Peter Schmidt, and Jaak Billiet. 2014. Measurement equivalence in cross-national research. *Annual Review of Sociology* 40:55-75. doi:10.1146/annurev-soc-071913-043137.

Davidov, Eldad, Peter Schmidt, and Shalom H. Schwartz. 2008. Bringing values back in: The adequacy of the European Social Survey to measure values in 20 countries. *Public Opinion Quarterly* 72:420-445. doi:10.1093/poq/nfn035.

Davidov, Eldad, and Pascal Siegers. 2010. Comparing basic human values in East and West Germany. In *Komparative empirische Sozialforschung* (Comparative empirical social research)*,* eds. Tilo Beckers, Klaus Birkelbach, Jörg Hagenah, and Ulrich Rosar, 43-63. Wiesbaden: VS Verlag.

De Beuckelaer, Alain, and Gilbert Swinnen. 2018. Biased latent variable mean comparisons due to measurement noninvariance: A simulation study. In *Cross-cultural research: Methods and applications*, *2nd edition,* eds. Eldad Davidov, Peter Schmidt, Jaak Billiet, and Bart Meuleman, 127-156. New York: Routledge Taylor & Francis Group.

Döring, Anna, Shalom H. Schwartz, Jan Cieciuch, Patrick J. F. Groenen, Valentina Glatzel, Justyna Harasimczuk, Nicole Janowicz, Maya Nyagolova, Rebecca E. Scheefer, Matthias Allritz, Taciano L. Milfont, and Wolfgang Bilsky. 2015. Cross-cultural evidence of value structures and priorities in childhood. *British Journal of Psychology* 106:675-699. doi:10.1111/bjop.12116.

Durkheim, Émile. 1897/1964. *Suicide*. Glencoe, IL: Free Press.

Guenole, Nigel. 2016. The importance of isomorphism for conclusions about homology: A Bayesian multilevel structural equation modeling approach with ordinal indicators. *Frontiers in Psychology* 7:289. doi:10.3389/fpsyg.2016.00289.

Hitlin, Steven, and Allyn Piliavin. 2004. Values: Reviving a dormant concept. *Annual Review of Sociology* 30:359-393. doi:10.1146/annurev.soc.30.012703.110640.

Hofstede, Geert. 2000. *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations, 2nd edition*. Beverly Hills, CA: Sage.

Horn, John L., and John J. McArdle. 1992. A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research* 18:117-144. doi:10.1080/03610739208253916.

Inglehart, Ronald, and Wayne E. Baker. 2000. Modernization, cultural change, and the persistence of traditional values. *American Sociological Review* 65:19–51.

Jak, Suzanne, Frans J. Oort, and Conor V. Dolan. 2013. A test for cluster bias: Detecting violations of measurement invariance across clusters in multilevel data. *Structural Equation Modeling* 20:265-282. doi:10.1080/10705511.2013.769392.

Jöreskog, Karl G. 1971. Simultaneous factor analysis in several populations. *Psychometrika* 36:409-426. doi:10.1007/BF02291366.

Kluckhohn, Clyde. 1951. Values and value-orientations in the theory of action: An exploration in definition and classification. In *Toward a general theory of action*, eds. Talcott Parsons and Edward A. Shils, 388–433. Cambridge, MA: Harvard University Press.

Knoppen, Desirée, and Willem Saris. 2009. Do we have combine values in the Schwartz' human values scale? A comment on the Davidov studies. *Survey Research Methods* 3: 91-103.

Lomazzi, Vera. 2018. Using alignment optimization to test the measurement invariance of gender role attitudes in 59 countries. *Methods, data, analyses: A journal for quantitative methods and survey methodology (mda)* 12(1): 77-103. doi:10.12758/mda.2017.09.

Magun, Vladimir, Maxim Rudnev, and Peter Schmidt. 2016. Within- and between-country value diversity in Europe: A typological approach. *European Sociological Review 32*:189-202. doi:10.1093/esr/jcv080.

Marsh, Herbert W., Jiesi Guo, Philip D. Parker, Benjamin Nagengast, Tihomir Asparouhov, Bengt O. Muthén, and Theresa Dicke. 2017. What to do when scalar invariance fails: The extended alignment method to multi-group factor analysis comparison of latent means across many groups. *Structural Equation Modeling*, in press. doi:10.1037/met0000113.

Meitinger, Katharina. 2017. Necessary but insufficient: Why measurement invariance tests need online probing as a complementary tool. *Public Opinion Quarterly* 81:447-472. https://doi.org/10.1093/poq/nfx009.

Merkle, Edgar C., and Yves Rosseel. 2016. blavaan: Bayesian structural equation modelling via parameter expansion. *arXiv: 1511.05604v2 [stat.CO]*. Retrieved from https://arxiv.org/abs/1511.05604 on June 4, 2018.

Millsap, Roger E. 2011. *Statistical approaches to measurement invariance*. New York: Routledge.

Munck, Ingrid, Carolyn Barner, and Judith Torney-Purta. 2017. Measurement invariance in comparing attitudes toward immigrants among youth across Europe in 1999 and 2009. The alignment method applied to IEA CIVED and ICCS. *Sociological Methods & Research*. doi:10.1177/0049124117729691

Muthén Bengt O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research* 22:376–398. doi:10.1177/0049124194022003006.

Muthén, Bengt O., and Tihomir Asparouhov. 2013. BSEM measurement invariance analysis. *Mplus Web Notes* 17:1-48.

Muthén, Bengt O., and Tihomir Asparouhov. 2014. IRT studies of many groups: The alignment method. *Frontiers in Psychology* 978:1-7. doi:10.3389/fpsyg.2014.00978.

Muthén, Bengt O., and Tihomir Asparouhov. 2017. Recent methods for the study of measurement invariance with many groups. Alignment and random effects. *Sociological Methods & Research* in press. doi:10.1177/0049124117701488.

Muthén, Linda K., and Bengt O. Muthén. 1998-2014. *Mplus user's guide.* Los Angeles: Muthén & Muthén.

Oberski, Daniel L. 2014. Evaluating sensitivity of parameters of interest to measurement invariance in latent variable models. *Political Analysis* 22:45-60. doi:10.1093/pan/mpt014.

Rokeach, Milton. 1973. *The nature of human values*. New York, NY: Free Press.

Rudnev, Maksim, Ekaterina Lytkina, Eldad Davidov, Peter Schmidt, and Andreas Zick. 2018. Testing measurement invariance for a second-order factor: A cross-national test of the alienation scale. *Methods, data, analyses: A journal for quantitative methods and survey methodology (mda)* 12(1):47-76. doi:10.12758/mda.2017.11

Rudnev, Maxim, Vladimir Magun, and Shalom Schwartz. 2018. Relations among higher order values around the world. *Journal of Cross-Cultural Psychology.* doi: 10.1177/0022022118782644

Ruelens, Anna, Bart Meuleman and Ides Nicaise. 2016. Examining measurement isomorphism of multilevel constructs: The case of political trust. *Social Indicators Research.* doi:10.1007/s11205-017-1799-6.

Schafer, Joseph L., and John W. Graham. 2002. Missing values: Our view of the state of the art. *Psychological Methods* 7(2):147-177. doi:10.1037//1082-989X.7.2.147.

Schwartz, Shalom H. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology*, *vol. 25*, ed. Mark Zanna, 1–65. London, UK: Academic Press.

Schwartz, Shalom H. 2003. A proposal for measuring value orientations across nations. In *Questionnaire development package of the European Social Survey*, 259-319. Retrieved from www.europeansocialsurvey.org, June 30, 2016.

Schwartz, Shalom H., and Jan Cieciuch. 2016. Values. In *The ITC international handbook of testing and assessment*, eds. Frederick T. L. Leong, Dave Bartram, Fanny M. Cheung, Kurt F. Geisinger, and Dragos Iliescu, 106-119. Oxford: Oxford University Press.

Schwartz, Shalom H., Jan Cieciuch, Michelle Vecchione, Eldad Davidov, Ronald Fischer, Constanze Beierlein, Alice Ramos, Markku Verkasalo, Jan-Erik Lönnqvist, Kursad Demirutku, Ozlem Dirilen-Gumus, and Mark Konty. 2012. Refining the theory of basic individual values. *Journal of Personality and Social Psychology* 103:663–688. doi:10.1037/a0029393.

Schwartz, Shalom H., Gila Melech, Arielle Lehmann, Steven Burgess, Mari Harris, and Vicki Owens. 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.

Sokolov, Boris. 2018. The index of emancipative values: Measurement model Misspecifications. *American Political Science Review* 112(*2*):395–408.

Steenkamp, Jan-Benedict E. M., and Hans Baumgartner. 1998. Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research* 25:78-90. doi:10.1086/209528

Steinmetz, Holger Rodrigo Isidor, Naissa Baeuerle. 2012. Testing the circular structure of human values: A meta-analytical structural equation modelling approach. *Survey Research Methods* 6 (1):61-75

Steinmetz, Holger. 2018. Estimation and comparison of latent means across cultures. In *Cross-cultural analysis: Methods and applications*, 2nd edition, eds. Eldad Davidov, Peter Schmidt, Jaak Billiet, and Bart Meuleman, 95-126. New York: Routledge Taylor & Francis Group.

Van de Schoot, Rens, Anouck Kluytmans, Lars Tummers, Peter Lugtig, Joop Hox, and Bengt O. Muthén. 2013. Facing off with Scylla and Charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology* 770:1-15. doi:10.3389/fpsyg.2013.00770.

Vandenberg, Robert J., and Charles E. Lance. 2000. A review and synthesis of the measurement invariance literature: Suggestions, practices and recommendations for organizational research. *Organizational Research Methods* 3:4-69. doi:10.1177/109442810031002.

Weber, Max. 1905/1958. *The protestant ethic and the spirit of capitalism*. New York: Scribner's.

Welzel, Christian, and Ronald F. Inglehart. 2016. Misconceptions of measurement equivalence: Time for a paradigm shift. *Comparative Political Studies* 49:1068-1094.

Zercher, Florian, Peter Schmidt, Jan Cieciuch, and Eldad Davidov. 2015. The comparability of the universalism value over time and across countries in the European Social Survey: Exact versus approximate measurement invariance. *Frontiers in Psychology* 733:1-11 doi:10.3389/fpsyg.2015.00733.

Table 1

The ten basic human values, four higher-order values, and the PVQ-21 items in the ESS (female version) to measure these values with their labels (The number next to each question item refers to the placement of that item in the PVQ-21 questionnaire.)

Item label	Items
1. Self-enhancement – Achievement	
ipshabt	4. It's important to her to show her abilities. She wants people to admire what she does.
ipsuces	13. Being very successful is important to her. She hopes people will recognize her achievements.
2. Self-enhancement – Power	
imprich	2. It is important to her to be rich. She wants to have a lot of money and expensive things.
iprspot	17. It is important to her to get respect from others. She wants people to do what she says.
3. Self-transcendence – Benevolence	
iphlppl	12. It's very important to her to help the people around her. She wants to care for their well-being.
iplylfr	18. It is important to her to be loyal to her friends. She wants to devote herself to people close to her.
4. Self-transcendence – Universalism	
ipeqopt	3. She thinks it is important that every person in the world should be treated equally. She believes everyone should have equal opportunities in life.
ipudrst	8. It is important to her to listen to people who are different from her. Even when she disagrees with them, she still wants to understand them.
impenv	19. She strongly believes that people should care for nature. Looking after the environment is important to her.
5. Conservation – Conformity	
ipfrule	7. She believes that people should do what they're told. She thinks people should follow rules at all times, even when no-one is watching.
ipbhprp	16. It is important to her always to behave properly. She wants to avoid doing anything people would say is wrong.
6. Conservation – Tradition	
ipmodst	9. It is important to her to be humble and modest. She tries not to draw attention to herself.
imptrad	20. Tradition is important to her. She tries to follow the customs handed down by her religion or her family.
7. Conservation – Security	
impsafe	5. It is important to her to live in secure surroundings. She avoids anything that might endanger her safety.
ipstrgv	14. It is important to her that the government ensures her safety against all threats. She wants the state to be strong so it can defend its citizens.
8. Openness – Self-direction	
ipcrtiv	1. Thinking up new ideas and being creative is important to her. She likes to do things in her own original way.
impfree	11. It is important to her to make her own decisions about what she does. She likes to be free and not depend on others.
9. Openness – Stimulation	
impdiff	6. She likes surprises and is always looking for new things to do. She thinks it is important to do lots of different things in life.
ipadvnt	15. She looks for adventures and likes to take risks. She wants to have an exciting life.
10. Openness – Hedonism	
ipgdtim	10. Having a good time is important to her. She likes to "spoil" herself.

impfun	21. She seeks every chance she can to have fun. It is important to her to do things that give her pleasure.

Table 2

Number of respondents included in the analysis for each round and country

	1st Round 2002-3	2nd Round 2004-5	3rd Round 2006-7	4th Round 2008-9	5th Round 2010-11	6th Round 2012-13	7th Round 2014-15
Belgium (BE)	1,819	1,734	1,767	1,704	1,674	1,809	1,720
Denmark (DK)	1,457	1,457	1,451	1,554	1,548	1,610	1,475
Finland (FI)	1,758	1,692	1,645	1,898	1,638	2,142	2,044
Germany (DE)	2,785	2,800	2,828	2,697	2,943	2,910	2,982
Hungary (HU)	1,564	1,407	1,409	1,388	1,404	1,919	1,460
Ireland (IE)	1,838	1,139	1,582	1,682	2,295	2,498	2,288
Netherlands (NL)	2,301	1,824	1,814	1,693	1,754	1,788	1,802
Norway (NO)	1,806	1,543	1,533	1,374	1,518	1,598	1,408
Poland (PL)	1,982	1,621	1,629	1,544	1,675	1,818	1,550
Portugal (PT)	1,417	1,987	2,117	2,220	2,035	2,062	1,209
Slovenia (SI)	1,390	1,297	1,329	1,172	1,238	1,159	1,113
Spain (ES)	1,638	1,544	1,802	2,520	1,862	1,820	1,857
Sweden (SE)	1,677	1,663	1,585	1,539	1,457	1,799	1,755
Switzerland (CH)	2,009	2,084	1,758	1,764	1,467	1,453	1,489
United Kingdom (GB)	1,748	1,806	2,301	2,230	2,315	2,212	2,176
Total	25,441	23,792	24,249	24,749	24,508	26,385	26,328

Note: Only countries that participated in all seven ESS rounds are included in the analysis.

Table 3

Model fit indices of the multiple-group confirmatory factor analyses across all countries and waves for each higher-order value (configural invariance model)

	$\chi 2$	df	CFI	RMSEA	SRMR
Self-enhancement	741.5	210	0.996	0.038 [0.035-0.041]	0.011
Self-transcendence	1907.0	315	0.989	0.053 [0.051-0.055]	0.015
Conservation	2758.2	630	0.988	0.043 [0.042-0.045]	0.017
Openness to change	4278.3	210	0.961	0.104 [0.101-0.107]	0.030

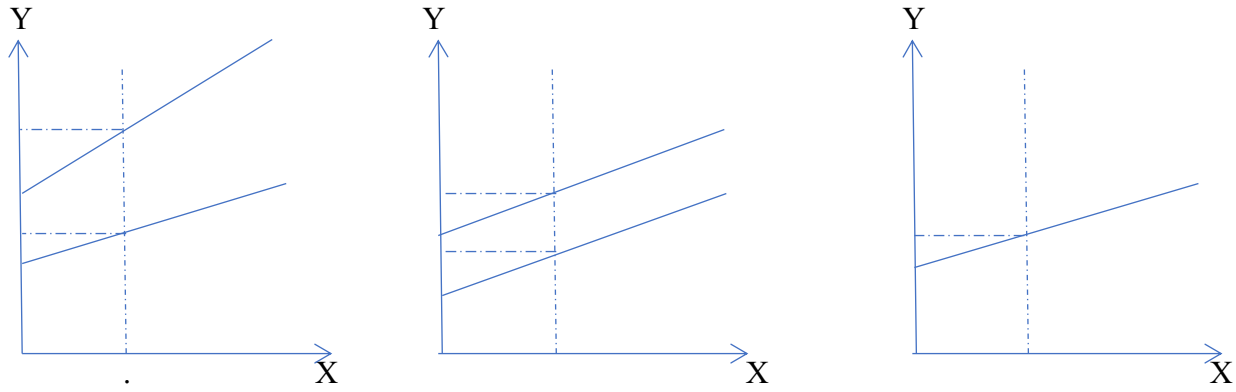
Note: df = degrees of freedom, CFI = Comparative Fit Index; RMSEA = Root Mean Square Error of Approximations; SRMR = Standardized Root Mean Square Residuals

Table 4

The number (and percentage) of noninvariant loadings and intercepts identified in the alignment optimization for each higher-order value

	Loadings	Intercepts	Average
Conservation	12% (73/630)	47% (294/630)	29% (367/1260)
Self-transcendence	12% (61/525)	45% (234/525)	28% (295/1050)
Openness (without hedonism)	14% (60/420)	54% (228/420)	34% (289/840)
Self-enhancement	6% (26/420)	59% (247/420)	33% (273/840)

Figure 1. Illustration of configural invariance, metric invariance, and scalar invariance across two countries



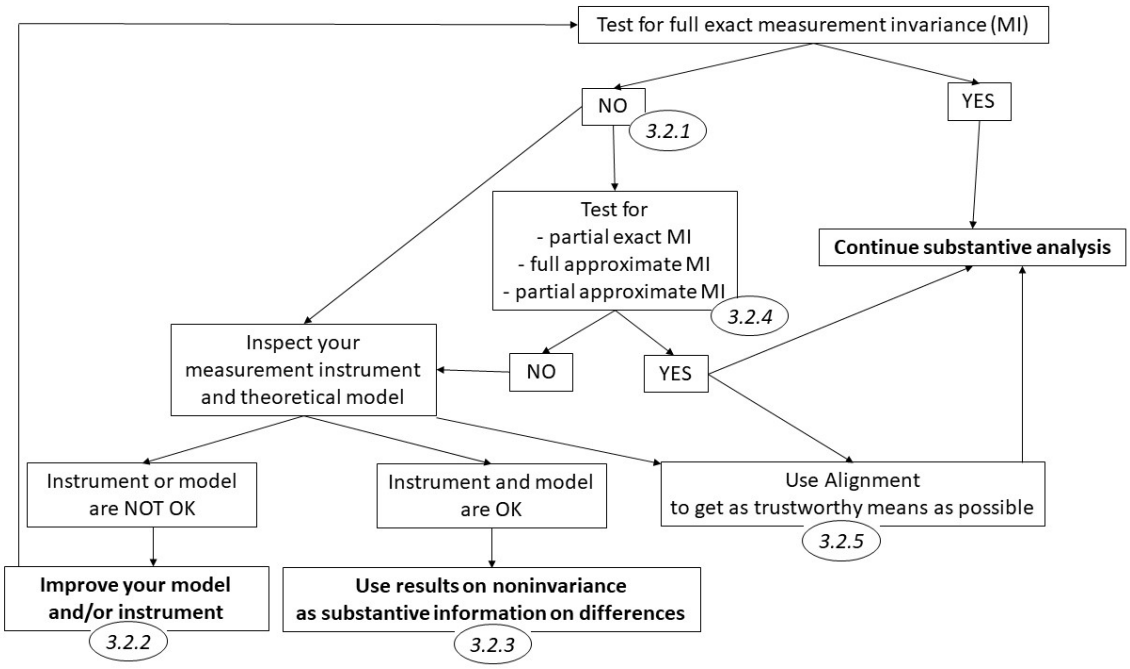
(a) Configural + Metric - Scalar -

(b) Configural + Metric + Scalar -

(c) Configural + Metric + Scalar +

Note: The X axis represents the latent variable mean; the Y axis represents the response to a survey question item measuring the latent variable. The diagonal represents the functional relation between the latent variable and the response to the survey question item in two countries (in unstandardized terms).

Figure 2. Overview of the procedures and decisions in measurement invariance testing when exact measurement invariance is not supported by the data.



Note: MI = measurement invariance; the numbers in the ellipses refer to subsections where we describe examples of a given procedure.

Figure 3. The circular motivational continuum of values (Source: Cieciuch et al. 2015)

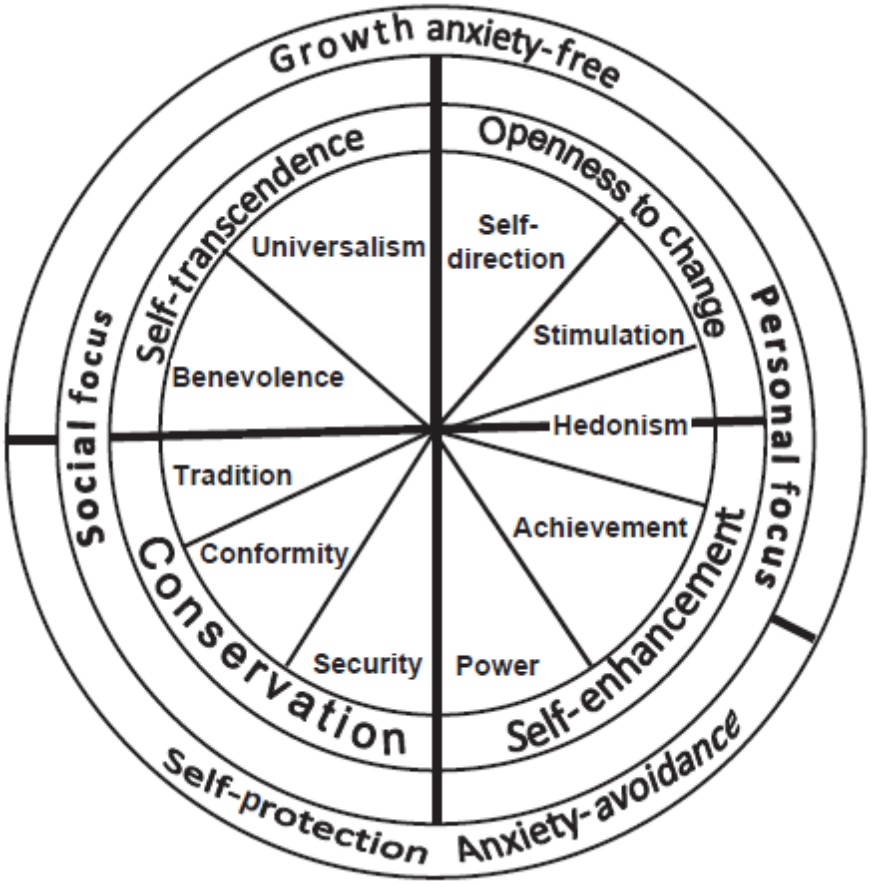
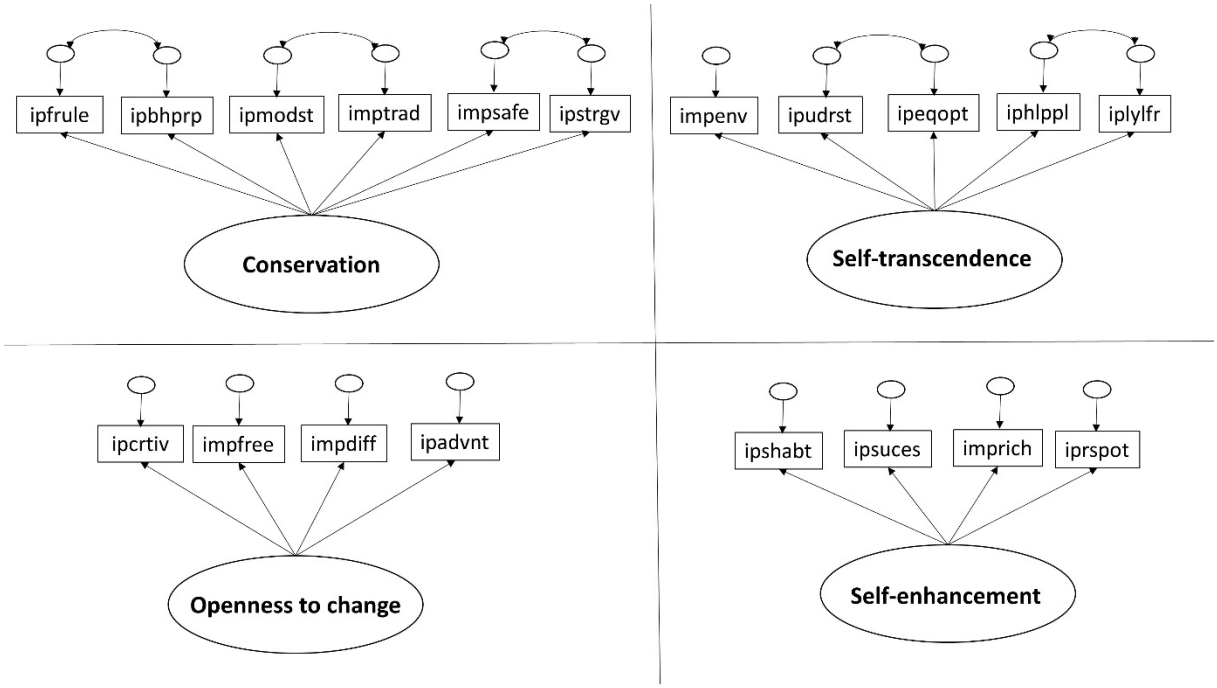


Figure 4. Measurement models for four higher-order values. Item abbreviations are presented in Table 1. Error correlations are allowed between items that were originally designed to measure the same single value in order to take into account their common variance.



Appendix

Table A1

Importance of conservation values: Country rankings across all ESS rounds (The value means estimated by the alignment optimization are in parentheses.)

Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
PL* (-0.413)	PL* (-0.455)	HU* (-0.47)	ES* (-0.314)	SI* (-0.447)	SI* (0)	SI* (-0.089)
ES* (-0.488)	HU* (-0.489)	ES* (-0.493)	HU* (-0.615)	HU* (-0.477)	PL* (-0.28)	PL* (-0.274)
HU* (-0.568)	IE* (-0.741)	PL* (-0.571)	PL* (-0.711)	PL* (-0.51)	ES* (-0.531)	HU* (-0.517)
IE* (-0.813)	ES* (-0.82)	SI* (-0.938)	SI* (-0.768)	ES* (-0.58)	HU* (-0.642)	ES* (-0.526)
SI* (-0.877)	SI* (-0.999)	IE* (-0.962)	IE* (-0.849)	IE* (-1.017)	IE* (-0.792)	IE* (-0.812)
PT* (-1.042)	BE (-1.164)	PT* (-1.016)	BE (-1.207)	GB (-1.081)	GB (-0.863)	CH (-1.086)
BE (-1.317)	GB (-1.184)	BE (-1.217)	GB (-1.302)	CH (-1.138)	CH (-1.008)	GB (-1.1)
FI (-1.345)	PT* (-1.266)	GB (-1.265)	PT* (-1.367)	DE* (-1.157)	BE (-1.052)	BE (-1.144)
GB (-1.348)	FI (-1.419)	DE* (-1.441)	DE* (-1.449)	PT* (-1.186)	PT* (-1.162)	PT* (-1.297)
DE* (-1.364)	DE* (-1.445)	FI (-1.463)	CH (-1.473)	BE (-1.318)	DE* (-1.17)	DE* (-1.303)
CH (-1.594)	CH (-1.494)	CH (-1.465)	FI (-1.503)	FI (-1.432)	FI (-1.426)	FI (-1.477)
NL (-1.631)	NL (-1.627)	NL (-1.721)	NL (-1.734)	DK (-1.601)	NL (-1.594)	DK (-1.644)
DK (-1.788)	DK (-1.845)	NO (-1.875)	NO (-1.998)	NL (-1.683)	DK (-1.699)	NO (-1.661)
NO (-2.154)	NO (-1.897)	DK (-2.009)	DK (-2.015)	NO (-1.942)	NO (-1.722)	NL (-1.688)
SE (-2.392)	SE (-2.312)	SE (-2.311)	SE (-2.31)	SE (-2.196)	SE (-1.928)	SE (-2.096)

*Note. * Country/round combinations where the number of noninvariant parameters exceeds 25%. Consequently, their means must be analyzed with caution.*

Table A2

Importance of self-transcendence values: Country rankings across all ESS rounds (The value means estimated by the alignment optimization are in parentheses.)

Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
CH* (-0.149)	CH* (0)	ES* (0.113)	ES* (0.223)	CH* (0.192)	ES* (0.318)	SI* (0.267)
ES* (-0.152)	ES* (-0.216)	CH* (0.082)	CH* (0.144)	ES* (0.185)	SI* (0.25)	ES* (0.205)
BE (-0.283)	IE (-0.255)	BE (-0.229)	IE (-0.13)	SI* (0.024)	CH* (0.127)	CH* (0.187)
DE* (-0.37)	BE (-0.289)	FI* (-0.404)	BE (-0.221)	DE* (-0.081)	DE* (0.035)	DE* (0.101)
FI* (-0.409)	HU* (-0.391)	DK* (-0.411)	DK* (-0.246)	DK* (-0.258)	FI* (-0.129)	SE (-0.064)
DK* (-0.433)	PL* (-0.454)	SI* (-0.42)	DE* (-0.247)	BE (-0.274)	SE (-0.137)	FI* (-0.095)
IE (-0.482)	DE* (-0.474)	PL* (-0.444)	SI* (-0.311)	HU* (-0.306)	DK* (-0.185)	DK* (-0.226)
PL* (-0.614)	FI* (-0.474)	GB (-0.449)	FI* (-0.355)	SE (-0.353)	BE (-0.188)	BE (-0.243)
SI* (-0.65)	DK* (-0.484)	DE* (-0.459)	GB (-0.431)	GB (-0.393)	GB (-0.2)	PL* (-0.261)
NL (-0.658)	GB (-0.488)	IE (-0.464)	NL (-0.513)	PL* (-0.399)	PL* (-0.206)	GB (-0.29)
HU* (-0.7)	NL (-0.58)	NL (-0.529)	PL* (-0.537)	NL (-0.484)	HU* (-0.276)	HU* (-0.372)
GB (-0.737)	SI* (-0.598)	HU* (-0.549)	HU* (-0.609)	FI* (-0.486)	IE (-0.284)	IE (-0.436)
PT* (-0.824)	NO (-0.905)	NO (-0.753)	SE (-0.698)	NO (-0.496)	NL (-0.436)	NO (-0.502)
SE (-0.99)	SE (-0.917)	PT* (-0.784)	NO (-0.782)	IE (-0.599)	NO (-0.441)	PT* (-0.568)
NO (-0.996)	PT* (-1.204)	SE (-0.865)	PT* (-1.136)	PT* (-0.976)	PT* (-0.935)	NL (-0.592)

*Note. * Country/round combinations where the number of noninvariant parameters exceeds 25%. Consequently, their means must be analyzed with caution.*

Table A3

Importance of self-enhancement values: Country rankings across all ESS (The value means estimated by the alignment optimization are in parentheses.)

Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
SI* (-0.101)	HU* (-0.114)	SI* (-0.025)	SI* (0.085)	SI* (0.257)	HU* (0.298)	SI* (0.337)
HU* (-0.163)	PL* (-0.186)	PL* (-0.177)	HU* (-0.113)	HU* (0.045)	SI* (0.286)	HU* (0.204)
PL* (-0.321)	SI* (-0.209)	HU* (-0.214)	PL* (-0.202)	PL* (-0.067)	PL* (0)	PL* (-0.191)
ES (-0.344)	PT* (-0.226)	PT* (-0.342)	BE (-0.41)	IE* (-0.202)	IE* (-0.149)	IE* (-0.218)
PT* (-0.454)	BE (-0.562)	BE (-0.469)	PT* (-0.42)	PT* (-0.238)	CH* (-0.17)	CH* (-0.273)
IE* (-0.503)	IE* (-0.562)	CH* (-0.548)	IE* (-0.425)	CH* (-0.345)	PT* (-0.22)	BE (-0.461)
BE (-0.603)	DE* (-0.604)	IE* (-0.563)	CH* (-0.449)	BE (-0.538)	BE (-0.379)	PT* (-0.503)
DE* (-0.621)	ES (-0.609)	DE* (-0.581)	DE* (-0.63)	GB (-0.553)	GB (-0.48)	DK* (-0.526)
GB (-0.669)	GB (-0.621)	NL* (-0.655)	NL* (-0.646)	DE* (-0.625)	DK* (-0.517)	NL* (-0.684)
CH* (-0.674)	CH* (-0.667)	GB (-0.658)	DK* (-0.697)	NL* (-0.63)	NL* (-0.557)	GB (-0.784)
DK* (-0.741)	DK* (-0.734)	ES (-0.716)	GB (-0.739)	DK* (-0.684)	DE* (-0.589)	DE* (-0.852)
NL* (-0.763)	NL* (-0.79)	DK* (-0.722)	NO* (-0.78)	ES (-0.801)	ES (-0.777)	NO* (-0.857)
NO* (-0.941)	NO* (-0.838)	NO* (-0.854)	ES (-0.787)	NO* (-0.887)	NO* (-0.777)	ES (-0.904)
SE (-1.012)	SE (-0.991)	SE (-1.022)	SE (-0.933)	FI* (-0.978)	SE (-0.81)	SE (-1.019)
FI* (-1.081)	FI* (-1.094)	FI* (-1.137)	FI* (-1.18)	SE (-1.085)	FI* (-1.152)	FI* (-1.216)

*Note. * Country/round combinations where the number of noninvariant parameters exceeds 25%. Consequently, their means must be analyzed with caution.*

Table A4

Importance of openness to change values (without hedonism): Country rankings across all ESS rounds (The value means estimated by the alignment optimization are in parentheses.)

Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
CH* (-0.33)	HU* (-0.364)	SI* (-0.337)	SI* (-0.333)	SI* (-0.147)	SI* (-0.189)	SI* (0)
SI* (-0.416)	CH* (-0.391)	NL* (-0.43)	IE* (-0.357)	CH* (-0.354)	CH* (-0.231)	CH* (-0.236)
DK* (-0.424)	SI* (-0.517)	CH* (-0.5)	CH* (-0.425)	HU* (-0.365)	NL* (-0.342)	DK* (-0.27)
HU* (-0.5)	NL* (-0.542)	HU* (-0.507)	NL* (-0.444)	NL* (-0.369)	HU* (-0.436)	DE* (-0.502)
ES* (-0.527)	BE (-0.613)	IE* (-0.643)	DK* (-0.567)	DK* (-0.488)	ES* (-0.452)	NL* (-0.505)
BE (-0.562)	DE* (-0.683)	DK* (-0.655)	HU* (-0.594)	DE* (-0.525)	IE* (-0.471)	HU* (-0.535)
NL* (-0.566)	FI (-0.683)	GB (-0.659)	BE (-0.648)	FI (-0.527)	DE* (-0.483)	FI (-0.561)
GB (-0.591)	DK* (-0.684)	DE* (-0.671)	GB (-0.659)	ES* (-0.533)	BE (-0.493)	BE (-0.576)
FI (-0.598)	IE* (-0.72)	BE (-0.674)	DE* (-0.662)	IE* (-0.56)	DK* (-0.499)	IE* (-0.59)
IE* (-0.62)	GB (-0.725)	ES* (-0.688)	ES* (-0.674)	GB (-0.662)	SE* (-0.52)	SE* (-0.6)
DE* (-0.633)	ES* (-0.772)	FI (-0.709)	FI (-0.71)	BE (-0.691)	GB (-0.534)	ES* (-0.628)
PL* (-0.752)	PL* (-0.774)	PL* (-0.778)	SE* (-0.791)	PL* (-0.754)	FI (-0.556)	GB (-0.678)
SE* (-1.017)	SE* (-0.961)	SE* (-0.94)	PL* (-0.804)	SE* (-0.82)	PL* (-0.667)	NO* (-0.753)
NO* (-1.021)	NO* (-0.996)	NO* (-0.991)	NO* (-0.879)	NO* (-0.872)	NO* (-0.73)	PL* (-0.84)
PT* (-1.027)	PT* (-1.511)	PT* (-1.198)	PT* (-1.23)	PT* (-1.011)	PT* (-1.016)	PT* (-0.921)

*Note. * Country/round combinations where the number of noninvariant parameters exceeds 25%. Consequently, their means must be analyzed with caution.*