Testing for Approximate Measurement Invariance of Human Values in the European Social Survey

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

Abstract: Measurement invariance is a necessary precondition for meaningful cross-country comparisons, and three levels have been differentiated: configural, metric, and scalar. Unfortunately, establishing the most stringent form, i.e., scalar measurement invariance, across groups is difficult. Recently, Muthén and Asparouhov proposed testing for approximate rather than exact measurement invariance as this may be sufficient for meaningful comparisons. Following their strategy, the results of cross-country approximate measurement invariance tests of the PVQ-21 scale to measure values in the European Social Survey (ESS) are presented (N = 274,447 respondents from 15 countries participating in all six rounds). Applying the new approximate method for the test of measurement invariance allows both using more moderate constraints of approximate equality of parameters across groups and exploring the extent of noninvariance. Approximate measurement invariance was established in almost all rounds for two higher-order values: openness to change and self-enhancement. In the case of the two other higher-order values, self-transcendence and conservation, approximate measurement invariance was established across a subset of countries.

DOI: https://doi.org/10.1177/0049124117701478

Accepted for a special issue on 'Measurement Invariance'

in Sociological Methods & Research

Testing for approximate measurement invariance of human values in the European Social Survey

Jan Cieciuch

University of Zurich, Switzerland

Eldad Davidov

University of Cologne, Germany,

and University of Zurich, Switzerland

René Algesheimer

University of Zurich, Switzerland

Peter Schmidt

University of Giessen, Germany

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

Testing for approximate measurement invariance of human values

in the European Social Survey

Abstract

Measurement invariance is a necessary precondition for meaningful cross-country comparisons, and three levels have been differentiated: configural, metric, and scalar. Unfortunately, establishing the most stringent form, i.e., scalar measurement invariance, across groups is difficult. Recently, Muthén and Asparouhov proposed testing for approximate rather than exact measurement invariance as this may be sufficient for meaningful comparisons. Following their strategy, the results of cross-country approximate measurement invariance tests of the PVQ-21 scale to measure values in the European Social Survey (ESS) are presented (N = 274,447 respondents from 15 countries participating in all six rounds). Applying the new approximate method for the test of measurement invariance allows both using more moderate constraints of approximate equality of parameters across groups and exploring the extent of noninvariance. Approximate measurement invariance was established in almost all rounds for two higher-order values: openness to change and self-enhancement. In the case of the two other higher-order values, self-transcendence and conservation, approximate measurement invariance was established across a subset of countries.

Key words:

Exact and approximate measurement invariance; Human values; European Social Survey; Bayesian analysis; PVQ-21

Introduction

A meaningful comparison of constructs across groups requires that the same construct is measured in the same way in each group; that is, the construct has to be measurement invariant. Methodologists suggest that measurement invariance should not be assumed but rather tested empirically (e.g. Millsap 2011). One particular scale has been subject to such rigorous measurement invariance tests: the 21-item Portrait Value Questionnaire (PVQ-21), which has been included in the European Social Survey (ESS) since its inception to measure human values. Indeed, this scale has been used in numerous cross-cultural studies, and it is still critical to determine whether it is adequate for use in comparative research.

Unfortunately, when testing for strict or scalar measurement invariance of the PVQ-21 scale across countries, researchers are often unable to establish it (Davidov et al. 2014). This result is unfortunate because it implies that cross-cultural research of human values and particularly mean comparisons may not be meaningful in such a case. At the same time, this result also raises the question of whether the criteria for the evaluation of measurement invariance are appropriate or too strict. Indeed, recent methodological developments suggest that measurement invariance tests are too strict and propose to alternatively test for approximate rather than exact measurement invariance (Muthén and Asparouhov 2013).

In the current study we subject the PVQ-21 scale to a test of its cross-country comparability. However, instead of testing for its exact measurement invariance properties as done in previous literature, we subject it to the more liberal approximate invariance test. Findings of approximate invariance may provide an insight into the extent of noninvariance across countries and may possibly permit using it in cross-cultural research more extensively, particularly to compare value means across European countries. We begin by presenting the concept of approximate measurement invariance and explain how it differs from the traditional exact measurement invariance test. Next, we briefly summarize previous (and

rather disappointing) results of exact measurement invariance testing of the PVQ-21 in the ESS. Finally, we present the results of approximate measurement invariance tests of the scale across 15 countries and in six ESS rounds (2002 – 2012) and discuss these findings.

Approximate versus exact measurement invariance

Davidov et al. (2014:58) defined measurement invariance as "a property of a measurement instrument… implying that the instrument measures the same concept in the same way across various subgroups of respondents" (p. 58). Horn and McArdle (1992) relate, in their seminal work, the issue of measurement invariance to the question of "whether or not, under different conditions of observing and studying phenomena, measurement operations yield measures of the same attribute" (p. 117).

Multiple-group confirmatory factor analysis (MGCFA) is the most common approach to test for measurement invariance (for other approaches see, e.g., Davidov, Schmidt, and Billiet 2011). Researchers typically differentiate between three levels of measurement invariance: configural, metric, and scalar. Configural invariance means that *the same* latent variables are measured by *the same* items in all groups. Metric invariance additionally requires that all loadings of items are *the same* across groups. Scalar invariance implies that both factor loadings and indicator intercepts are *the same* across groups (Vandenberg and Lance 2000).

The key conclusion of whether measurement invariance at a specific level exists can be drawn based on a comparison of the model fit at a given level of invariance against the model fit of a less constrained model (Chen 2007). Thus, metric invariance is established when the fit of a model with all loadings constrained to be the same across groups is not considerably worse than that of a model without these constraints (at the configural level). Analogously, scalar measurement invariance is established when the fit of a model with all

loadings and intercepts constrained to be the same across groups does not deteriorate considerably compared to a model at the metric level (where only factor loadings are constrained to be equal across all groups). Chen (2007) provides cut-off criteria to determine whether a model becomes considerably worse or not.

In cases where measurement invariance is not established, some researchers have proposed testing for partial (metric or scalar) measurement invariance. They suggest that partial invariance is sufficient for performing meaningful comparisons (Byrne, Shavelson, and Muthén 1989; Steenkamp and Baumgartner 1998). Partial invariance is supported when the parameters of at least two indicators (loadings at the metric level and loadings plus intercepts at the scalar level of measurement invariance) are equal across groups.

The idea of establishing approximate (rather than exact or partial) measurement invariance was proposed recently by Muthén and Asparouhov (2013). It replaces the key requirement of *equality* of parameters across groups in the so-called traditional exact measurement invariance approach with the requirement that parameters are *approximately equal* (Davidov et al. 2015a; Muthén and Asparouhov, 2013; van de Schoot et al. 2013). This more liberal test of approximate invariance is performed within a Bayesian framework. In this framework, the difference in a parameter across groups is treated as a variable with a predefined distribution. For example, a distribution which requires both the *mean* and the *variance* of the difference of a parameter across groups to be zero reflects exact measurement invariance. A distribution may also define a (very) large variance for the parameter difference. In such a case, it reflects noninvariance for that parameter. Approximate measurement invariance is situated in-between these two extremes. It typically requires that the mean difference of a parameter is zero and the variance of the parameter difference is small. The size of the variance reflects the level of approximation: The smaller the variance of the

difference, the more restrictive the model is and the more similar it is to an exact measurement invariance model.

A detailed explanation of the Bayesian approach for testing approximate measurement invariance is beyond the scope of this study and can be found in the papers of Muthén and Asparouhov (2013, 2016 in this volume) and Van de Schoot et al. (2013). One of the advantages of the Bayesian approach is that researchers can introduce previous knowledge into the analysis. If the researcher knows what size of differences between parameters does not significantly bias substantive conclusions, this knowledge can be introduced into the analysis by determining the size of the variance of the parameter differences in the prior distribution. For the more realistic situation in which previous knowledge is absent, Van de Schoot et al. (2013) provide recommendations about which variance may be chosen for the priors. They base their recommendations on simulation studies, which demonstrate that substantive conclusions are not biased when parameters (factor loadings and intercepts) vary across groups to a certain extent. Their simulations suggest that a variance of .05 for the parameter differences may be small enough to not distort the substantive results (e.g., of latent mean comparisons) and large enough to make the assumption realistic.

The test for approximate measurement invariance, implemented in the Mplus software package (Muthén and Muthén 1998-2014), provides researchers with two types of output to evaluate the model fit. The first type corresponds to the global fit measures in the exact approach (Bollen 1989) and includes the posterior predictive *p* value (ppp) and the credibility interval (CI) for the difference between the observed and the replicated chi-square scores. According to Muthén and Asparouhov (2013) and Van de Schoot et al. (2013), the Bayesian model fits the data well when the ppp is not significant and the CI contains zero. The second type of output obtained in Mplus to evaluate the fit of the model is the so-called difference output. This part of the output lists all the noninvariant parameters in each group. Based on

this list, researchers may conclude which countries and/or items are approximately invariant and which ones are not.

The human values theory and the PVQ-21 measurement of values in the ESS

Schwartz (1992, 2003; Schwartz et al. 2012) defines values as broad, transsituational goals that vary in importance and serve as guiding principles in the life of a person or group. According to Schwartz's theory, values are organized around a circle with close values sharing a similar underlying motivation and values further apart or opposite of each other having contrasting underlying motivations. Values are key for describing both individuals (Schwartz 1992) and societies (Schwartz 2006) and for explaining behavior, attitudes, and opinions on the individual level or social change on the group level (Schwartz 2006). This is why the designers of the ESS decided to include, since its inception, a measurement of values in each of its rounds, the 21-item Portrait Value Questionnaire (PVQ-21) developed by Schwartz (2003). The PVQ-21 measures 10 basic values: achievement, power, hedonism, self-direction, stimulation, universalism, benevolence, conformity, tradition, and security. According to the theory, the 10 values form four higher-order values: self-enhancement which opposes self-transcendence, and conservation which opposes openness to change. The values and their higher-order dimensions are presented in Table 1.

Table 1 about here

Measurement invariance tests of human values: Previous findings and the current study

Several studies tested the measurement invariance properties of the PVQ-21 scale using ESS data (Davidov 2008, 2010; Davidov, Schmidt, and Schwartz 2008). All of these studies led to the following two conclusions: First, only 7 out of 10 values could be identified. Second, only

metric invariance could be established for these seven values, but not scalar invariance. Below we discuss these results in more detail and present the goals of the current study.

Davidov et al. (2008) showed that it is necessary to unify some pairs of adjacent values (power with achievement, benevolence with universalism, conformity with tradition). Unifying pairs of neighboring values does not contradict the main assumption of the theory, which postulates that the continuum of values can be divided in various ways, although it runs counter to expectations because the PVQ-21 was developed to measure 10 values. Knoppen and Saris (2009) proposed an explanation of why it is necessary to unify some values. They argued that the unification is a consequence of the circular structure of values and the choice strategy of items in the ESS. The aims of the strategy were twofold and somewhat contradictory: The first aim was to include in the ESS as few indicators as possible due to economic reasons; the second aim was to substantively cover the entire circle of values. Consequently, the items developed to measure the same value were not homogenous enough, and the items developed to measure neighboring values were not different enough. Therefore, it was not possible to establish sufficient levels of discriminant and convergent validity for the value measurements (see also Beierlein et al. 2012). Moreover, Knoppen and Saris (2009) proposed to use separate models at a time for each pair of adjacent values in the circle (rather than the whole model) when testing their validity properties. We followed this strategy and ran models for each higher-order value separately for the following reasons: First, we were interested in testing for approximate invariance across countries rather than in testing the theoretically postulated circular structure of the values. Second, positive cross-loadings on neighboring values and negative cross-loadings on opposite values are inherent in the theory. However, they may make the model quite complex and, thus, add complexity which is not directly relevant to the main goal of the current study – assessing cross-group comparability. Third, in the meantime, running models for each higher-order value separately has become a

common practice in the value literature (see, e.g., Cieciuch et al. 2014; Cieciuch and Schwartz 2012; Schwartz and Butenko 2014).

In the current study we address the problem of scalar (exact) noninvariance (Davidov et al., 2008) by subjecting the values to approximate measurement invariance tests. It may well be the case that previous studies failed to establish scalar invariance for the values due to the excessively strict criteria of the exact approach. Indeed, a study by Cieciuch and colleagues (2014), which compared the results of exact and approximate measurement invariance tests of 19 values differentiated in the refined value theory (Schwartz et al., 2012) and used nonrepresentative, mostly student samples, reported considerably higher levels of measurement invariance in the approximate compared to the exact approach. The study of Cieciuch at al. (2014) was innovative, but it did not use population samples. In addition, data used in the study were collected in only eight countries. This limitation was addressed by a recent study conducted by Zercher et al. (2015) which tested for approximate measurement invariance simultaneously across countries and time points using the ESS data. However, this study focused on a measurement invariance test of only one value – universalism. Thus, its goal was rather to demonstrate how such a test may be conducted on a very large number of groups. Our study is much more conclusive. Specifically, we address the above limitations by systematically testing for measurement invariance of the measurements of all human values across a large set of countries and time points and using population samples. In addition, and complementary to previous research, we also conduct sensitivity analyses with three different sizes of prior variances for the differences between the loadings and intercepts across countries. We conducted separate approximate measurement invariance tests on values measured in the first six rounds of the ESS data across all countries that participated in all rounds. This is, to the best of our knowledge, the most extensive approximate invariance test in the literature so far.

Data, measurements, and analytical approach

In the analysis we included the 15 countries that participated in each of the first six ESS rounds (2002/3, 2004/5, 2006/7, 2008/9, 2010/11, and 2012/13). The data for these countries were retrieved from the ESS website (www.europeansocialsurvey.org).[1] We followed recommendations on the ESS website and included, in our analysis, only respondents with no more than 5 missing values and no more than 16 identical responses for the 21 value items. Thus, we analyzed data from 274,447 respondents. Table 2 summarizes the number of respondents for each round and country included in the analysis.

Table 2 about here

The PVQ-21 items developed by Schwartz (2003) are presented in Table 1. Each item consists of two sentences describing a portrait from a male or female perspective. The portraits contain goals, aspirations, or desires that point implicitly to the importance of a value. For each item, the respondents answer the question "How much like you is this person?" on a scale ranging from 1 (not like me at all) to 6 (very much like me).

Analytical method

We began by evaluating the model fit of each higher-order value (i.e., conservation, openness to change, self-transcendence, and self-enhancement) across the 15 ESS countries and in each round separately using multiple-group confirmatory factor analysis (MGCFA). To capture the fact that specific pairs of items loading on a given higher-order value factor were originally developed to measure a particular basic value (and may, therefore, be more similar to each

1 Further information on data collection procedures, the full questionnaire, response rates, and methodological documentation are available on the ESS website.

other than to the other items measuring the same higher-order value), we released the correlations between the measurement errors of these item pairs for all the models presented below. For example, conservation was measured by six items including two conformity items, two tradition items, and two security items, thus, the error correlation of the two conformity items, the error correlation of the two tradition items, and the error correlation of the two security items were freed. In the case of self-transcendence, we freed the correlation between the two benevolence items and the errors of two (out of three) items measuring universalism (the importance of equality and tolerance), because the third universalism item measuring the importance of preserving nature involved a somewhat different facet of universalism.

Hedonism is located between openness to change and self-enhancement. Several studies consider it to be part of the openness to change higher-order value (see, e.g., Davidov et al. 2008). However, including hedonism to any higher-order value may be problematic, because the theory indicates that it is located between two higher-order values openness to change and self-enhancement, rather than belonging to either of them. Thus, we ran our models for openness to change twice, that is, with and without the hedonism value.[2]

We evaluated the model fit based on the cut-off criteria proposed in the literature by Hu and Bentler (1999) and Marsh, Hau, and Wen (2004) and regarded root mean square error of approximations (RMSEA) and standardized root mean square residuals (SRMR) smaller than .08 and comparative fit indices larger than .90 as indications of an acceptable model fit. Acceptable model fit allows us to consider each of the four higher-order latent variables as meaningful representations of our four higher-order values that may be used in further

2 There was one exception in which we had to constrain one error correlation to zero because of estimation problems: In the exact (but not in the approximate) test of openness without the hedonism value. This model had four items and one latent variable. The program allowed releasing the error correlation of one item pair only. In this model we released the correlation between the two stimulation items (and not between the two self-direction items). This decision was based on the refined value theory (Schwartz et al. 2012): The latter pair refers to two facets of self-direction in the new value theory, namely, action and thought, while the former pair refers to only one value in the refined theory, namely, stimulation. In the approximate tests there were no exceptions and all corresponding error correlations were released according to the logic just outlined (i.e., between the errors of those items originally developed to measure a particular basic value).

analysis. We did not test for invariance at this stage. Next, we tested for approximate measurement invariance of each of the four higher-order values.

In the approximate invariance tests we set the prior means of the differences between loadings and intercepts across countries to zero and the prior of the variance to .05. Then, we ran additional robustness checks in which we also used a more restrictive prior of .01 and a less restrictive prior of .1 for each higher-order value and in each ESS round. We estimated the correlations between the country rankings derived from the value latent means in each of these analyses to determine how sensitive the value mean rankings of the countries were to the choice of the priors. A very high correlation between country mean rankings based on different priors would indicate that the prior choice did not significantly influence our substantive results (see also Meuleman 2012; Oberski 2014).

The evaluation of the model was based on the ppp and the 95% CI. Nonsignificant ppp and a 95% CI that contained zero were treated as indications of an acceptable model fit (Muthén and Asparouhov 2013; Van de Schoot et al. 2013).

If approximate invariance could not be established across all 15 countries, we inspected the difference output (provided by the software package Mplus). In the difference output we we determined, for each higher-order value, those items with loadings and/or intercepts that are significantly noninvariant (i.e., deviating significantly from the average parameter across groups). Based on these results, a subset of countries with only small deviations from the average parameter could be identified in each round. Models for higher-order values were run on these particular subsets of countries. By using this procedure we tried to identify ,in a rather explorative way, the largest country subsets. However, also other smaller subsets of invariant countries may be selected. All analyses were conducted using the software package Mplus 7.3 (Muthén and Muthén 1998-2014).

Results

Table 3 presents the global fit measures of the MGCFA analyses (configural invariance models without constraints) for each higher-order value in each round separately.[3] Models for conservation, self-transcendence, and self-enhancement fit the data well. The RMSEA indices for the openness to change value that included the hedonism items were somewhat above the cut-off criteria in all rounds, so we reanalyzed the openness to change models without the hedonism items. This model fit the data very well. Furthermore, all factor loadings were considerable and exceeded 0.4 for all items (Brown 2015).

Table 3 about here

Assessing approximate measurement invariance across 15 countries for each higher-order value and in each round

Table 4 presents the global fit measures for the approximate measurement invariance test across 15 countries of each higher-order value and in each ESS round separately with a prior variance that equals .05 (the results for the 0.1 and 0.01 prior variances are presented in Appendix A). It turned out that models for self-enhancement and openness to change (without hedonism) fit the data quite well in almost all rounds (with three boundary cases: self-enhancement in the sixth Round and openness to change in the first and fourth Rounds). In other words, the two higher-order values displayed approximate measurement invariance. However, the models for self-transcendence and conservation displayed a rather poor fit to the data.[4] Choosing a more liberal prior of 0.1 did not improve the fit sufficiently. In agreement with previous findings (e.g., Davidov et al. 2008), the loadings of all items were usually approximately invariant across most countries in all rounds and for all higher-order values.

3 We enclose an Appendix with detailed results. Further details about the models and results may be available from the first author upon request.
4 The full output is available from the first author upon request.

However, results showed that most intercepts in the higher-order values self-transcendence and conservation were noninvariant.

Table 4 about here

Approximate measurement invariance tests for self-transcendence and conservation across subsets of countries

Whereas the approximate invariance test failed to establish measurement invariance for self-transcendence and conservation across all 15 countries, it could very well be the case that a subset of countries would display approximate measurement invariance. The Mplus difference output provides information about all the deviations of factor loadings and intercepts for all items in all countries and rounds. Information from this output is provided in Appendix B. This output can guide us in the selection of subsets of countries in which approximate invariance can be established for each round by allowing us to specifically identify a subset of countries with no significant deviation of loadings and/or intercepts from the average. Depending on the research goals and the analyses performed, it can be done for analyses on a given round or on all rounds simultaneously. The output presented in Appendix B is based on the analyses using priors of .05. Different priors may lead to identifying a slightly different subset of invariant countries.

For self-transcendence and conservation values we selected, from this output, those countries in each round whose intercepts differed least from the average, and we ran an approximate invariance test on these countries to confirm this observation. Finally, after confirming full or partial scalar approximate invariance for these subgroups of countries, we listed them in Table 5.

Table 5 about here

Conservation was approximately invariant across 10 countries (Belgium, Finland, Germany, Ireland, Netherlands, Poland, Portugal, Switzerland, United Kingdom, and Slovenia) in all six ESS rounds. Self-transcendence was approximately invariant across the 12 countries (Belgium, Finland, Germany, Ireland, Netherlands, Norway, Poland, Portugal, Spain, Sweden, Switzerland, and United Kingdom) in all six ESS rounds.

In order to test the robustness of our results we estimated the correlations between the country mean rankings for each higher-order value and round derived from the latent means estimated in models with different priors. Results remained intact and country mean rankings remained practically stable (correlations of the mean rankings varied between .983 and 1.000). All correlations and latent means obtained in models with different priors are presented in Appendix A.

Discussion

Measurement invariance is a necessary precondition for all meaningful comparisons across groups. In particular, scalar invariance is a precondition for meaningful mean comparisons across groups. The traditional approach usually applied to test for measurement invariance (the so-called exact measurement invariance approach) often leads to the rejection of scalar measurement invariance. As a consequence, all comparisons employing observed means, composite means, and latent means may be biased (Millsap 2011, Steinmetz 2011). Previous measurement invariance tests of the PVQ-21 scale, a human values measure which has been included in all ESS rounds and which has been extensively used by diverse researchers, also failed to demonstrate full or even partial scalar invariance properties. This was particularly unfortunate because, based on this result, cross-cultural researchers who were interested in comparing value scores across individuals and groups could not do so in a meaningful way.

Muthén and Asparouhov (2013) recently proposed to test for approximate rather than exact measurement invariance. They suggested that traditional exact measurement invariance tests may be too strict and that approximate measurement invariance tests may conclude that concepts are after all comparable. In the current study we subjected the PVQ-21 scale included in the ESS to a repeated and rigorous test of measurement invariance in the first six ESS rounds. However, this time we tested for approximate rather than exact measurement invariance and examined it for those countries that participated in all of the first six ESS rounds.

Our results revealed that approximate invariance of openness to change and self-enhancement was established across all countries in all ESS rounds, and subsets of countries could be identified where full or partial approximate scalar invariance was supported for self-transcendence and conservation. Thus, we could determine two values, self-enhancement and openness to change, where approximate scalar invariance was supported by the data and where comparisons of mean scores may be meaningfully conducted across all countries. In addition, we determined two values, self-transcendence and conservation, whose means may be compared across quite a large subset of countries.

Our study is not free of limitations. We conducted all analyses under the assumption that .05 variance of the difference in parameters is small enough not to distort substantive conclusions. We based our assumption on simulation studies conducted by Muthén and Asparouhov (2013) and Van de Schoot et al. (2013). However, the study of Van de Schoot indicated that a prior of .05 may lead to biased conclusions in specific circumstances. Thus, in order to test the robustness of our results, we ran additional models with a more restrictive prior (0.01) and a less restrictive prior (0.1). Our conclusions were essentially the same, and country mean ranking remained stable with different priors. Nevertheless, further simulation studies are needed to establish fixed criteria for the selection of the size of the prior variance.

Another issue still awaiting further research are the criteria for evaluating approximate measurement invariance models. We based our model evaluations on the criteria proposed by Muthén and Asparouhov (2013) and Van de Schoot et al. (2013), but further evaluation criteria would be useful for allowing a more precise determination of the fit of models assessing approximate measurement invariance to the data (see also Meuleman 2012; Oberski 2014; Kuha and Moustaki 2015). Furthermore, future studies should address in more detail the robustness of the global fit measures ppp and CI to sample size. Finally, although we suggested that specific countries are equivalent, we did not try to systematically explain why specific countries were not invariant for two of the four higher-order values. At first sight, there seems to be little in common among countries which turned out to be approximately invariant. Indeed, both methodological and substantive reasons could affect the item functioning of specific items differently across countries and time points (Davidov et al. 2012, 2016). Future studies may try to provide explanations for such patterns. Finally, an even more rigorous test could examine invariance properties across all time points and countries simultaneously. This would imply a particularly large number of groups included in the test, and it may lead to computational problems (Zercher et al. 2015). However, such a test may be necessary, if one wishes to conduct mean comparisons meaningfully both across countries and time points.

In sum, our findings may stimulate further comparative substantive research on values using ESS data. Such research has been, thus far, hindered by the results obtained by Davidov et al. (2008) regarding the lack of scalar measurement invariance of values. Establishing approximate invariance provides an opportunity to bridge the gap allowing value research to move forward. In our analyses, approximate invariance was established for two higher-order values across all countries and for the two remaining higher-order values across most countries. These findings are somewhat more promising than those obtained using the stricter

exact measurement invariance procedure in previous studies. In fact, the current findings challenge, at least to some extent, the previous rather disappointing conclusions on the cross-country invariance properties of human values as measured in the European Social Survey. Furthermore, the current findings suggest that it may be insightful to use the method and test for approximate measurement invariance in different national and international large data-generating programs also for other relevant theoretical constructs measured by multiple indicators such as trust, subjective well-being or various attitudes for which exact measurement invariance has been rejected in previous studies. Indeed, approximate measurement invariance test may succeed to establish invariance where exact tests have failed to do so.

References

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(1): 25–36.

Brown Timothy A. 2015. *Confirmatory Factor Analysis for Applied Research. Second Edition.* New York: Guilford Press.

Bollen, Kenneth A. 1989. *Structural Equations with Latent Variables.* New York: Wiley.

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(3):456-66.

Chen, Fang Fang. 2007. "Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance." *Structural Equation Modeling* 14(3):464-504. doi:10.1080/10705510701301834.

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 Scale to Measure 19 Human Values." *Frontiers in Psychology* 5:982. doi:10.3389/fpsyg.2014.00982.

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(3): 321-28. doi:10.1080/00223891.2012.655817.

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. 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* 51:171-91. doi:10.1177/0020715210363534.

Davidov, Eldad, Jan Cieciuch, Bart Meuleman, Peter Schmidt, René Algesheimer, and Mirjam Hausherr. 2015a. “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, Jan Cieciuch, Anabel Kuntz, Daniel Seddig, and Peter Schmidt. 2016. “Explaining Measurement Non-Equivalence Using Multilevel Structural Equation Modeling”. (Submitted to this volume).

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: 558-75. doi:10.1177/0022022112438397.

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 Jaak Billiet. eds. 2011. *Cross-Cultural Analysis: Methods and Applications.* New York: Routledge.

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* 72:420-45. doi:10.1093/poq/nfn035.

Horn, John L. and J. J. McArdle. 1992. “A Practical and Theoretical Guide to Measurement Invariance in Aging Research.” *Experimental Aging Research* 18(3):117-44. doi:10.1080/03610739208253916.

Hu, Litze and Peter M. Bentler. 1999. “Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria versus New Alternatives.” Structural Equation Modeling 6(1):1-55. doi:10.1080/10705519909540118.

Kuha, Jouni and Irini Moustaki. 2015. “Nonequivalence of Measurement in Latent Variable Modeling of Multigroup Data: A Sensitivity Analysis.” Psychological Methods. Advance online publication (July 27). http://dx.doi.org/10.1037/met0000031.

Knoppen, Desirée and Willem Saris. 2009. “Schwartz’ Theory of Human Values: Balancing Homogeneity of Reflective Items and Theoretical Coverage.” *RECSM Working Paper, 9*: 1-40. Barcelona: University Pompeu Fabra.

Marsh, Herbert W., Kit-Tai Hau, and Zhonglin Wen. 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(3):320-41. doi:10.1207/s15328007sem1103_2.

Meuleman Bart. 2012. “When are Intercept Differences Substantively Relevant in Measurement Invariance Testing?” Pp. 97–104 in *Methods, Theories, and Empirical Applications in the Social Sciences: Festschrift for Peter Schmidt*, edited by Samuel Salzborn, Eldad Davidov, and Jost Reinecke. Heidelberg: Springer VS.

Millsap, Roger E. 2011. *Statistical Approaches to Measurement Invariance*. New York: Routledge.

Muthén, Bengt O. and Tihomir Asparouhov. 2013. *BSEM Measurement Invariance Analysis*. Mplus Web Notes, 17.

Muthén, Bengt O. and Tihomir Asparouhov. 2016. “New Methods for the Study of Measurement Invariance with Many Groups” (submitted to this volume).

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

Schwartz, Shalom H. 1992. “Universals in the Content and Structure of Values: Theory and Empirical Tests in 20 Countries.” Pp. 1-65 in *Advances in Experimental Social Psychology*, Vol. 25, edited by Mark P. Zanna. New York: Academic Press.

Schwartz, Shalom H. 2003. “A Proposal for Measuring Value Orientations across Nations.” Pp. 259-319 in *Questionnaire Development Package of the European Social Survey*. Retrieved from www.europeansocialsurvey.org.

Schwartz, Shalom H. 2006. “A Theory of Cultural Value Orientations: Explication and Applications.” *Comparative Sociology* 5(2-3): 137-182. doi:10.1163/156913306778667357.

Schwartz, Shalom H., and Butenko, Tania. 2014. “Values and Behavior: Validating the Refined Value Theory in Russia.” *European Journal of Social Psychology* 44: 719-813. doi:10.1002/ejsp.2053.

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(4):663-88. doi:10.1037/a0029393.

Steenkamp, Jan-Benedict E.M. and Hans Baumgartner. 1998. “Assessing Measurement Invariance in Cross-National Consumer Research.” *Journal of Consumer Research* 25(1):78-90. doi:10.1086/209528.

Steinmetz, Holger. 2011. “Analyzing Observed Composite Differences Across Groups: Is Partial Measurement Invariance Enough?” *Methodology* 9, 1-12. doi: 10.1027/1614-2241/a000049.

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* 4:770. 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(1):4-70. doi:10.1177/109442810031002.

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* 6:733. doi:10.3389/fpsyg.2015.00733.

Table 1

The 10 basic human values, four higher-order values, and the PVQ-21 items in the ESS (male version) to measure these values with their labels (the number before 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 him to show his abilities. He wants people to admire what he does.
ipsuces	13. Being very successful is important to him. He hopes people will recognize his achievements.
2. Self-enhancement – Power	
imprich	2. It is important to him to be rich. He wants to have a lot of money and expensive things.
iprspot	17. It is important to him to get respect from others. He wants people to do what he says.
3. Self-transcendence – Benevolence	
iphlppl	12. It's very important to him to help the people around him. He wants to care for their well-being.
iplylfr	18. It is important to him to be loyal to his friends. He wants to devote himself to people close to him.
4. Self-transcendence – Universalism	
ipeqopt	3. He thinks it is important that every person in the world should be treated equally. He believes everyone should have equal opportunities in life.
ipudrst	8. 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.
impenv	19. He strongly believes that people should care for nature. Looking after the environment is important to him.
5. Conservation – Conformity	
ipfrule	7. He believes that people should do what they're told. He thinks people should follow rules at all times, even when no-one is watching.
ipbhprp	16. It is important to him always to behave properly. He wants to avoid doing anything people would say is wrong.
6. Conservation – Tradition	
ipmodst	9. It is important to him to be humble and modest. He tries not to draw attention to himself.
imptrad	20. Tradition is important to him. He tries to follow the customs handed down by his religion or his family.
7. Conservation - Security	
impsafe	5. It is important to him to live in secure surroundings. He avoids anything that might endanger his safety.
ipstrgv	14. It is important to him that the government ensures his safety against all threats. He 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 him. He likes to do things in his own original way.
impfree	11. It is important to him to make his own decisions about what he does. He likes to be free and not depend on others.
9. Openness – Stimulation	
impdiff	6. He likes surprises and is always looking for new things to do. He thinks it is important to do lots of different things in life.
ipadvnt	15. He looks for adventures and likes to take risks. He wants to have an exciting life.
10. Openness – Hedonism	
ipgdtim	10. Having a good time is important to him. He likes to “spoil” himself.
impfun	21. He seeks every chance he can to have fun. It is important to him to do things that give him pleasure.

Table 2

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

	1st Round	2nd Round	3rd Round	4th Round	5th Round	6th Round
	2002-3	2004-5	2006-7	2008-9	2010-11	2012-13
1. Belgium	1,819	1,734	1,767	1,704	1,674	1,809
2. Denmark	1,457	1,457	1,451	1,554	1,548	1,610
3. Finland	1,758	1,692	1,645	1,898	1,638	2,142
4. Germany	2,785	2,800	2,828	2,697	2,943	2,910
5. Hungary	1,564	1,407	1,409	1,388	1,404	1,919
6. Ireland	1,838	1,139	1,582	1,682	2,295	2,498
7. Netherlands	2,301	1,824	1,814	1,693	1,754	1,788
8. Norway	1,806	1,543	1,533	1,374	1,518	1,598
9. Poland	1,982	1,621	1,629	1,544	1,675	1,818
10. Portugal	1,417	1,987	2,117	2,220	2,035	2,062
11. Slovenia	1,390	1,297	1,329	1,172	1,238	1,159
12. Spain	1,638	1,544	1,802	2,520	1,862	1,820
13. Sweden	1,677	1,663	1,585	1,539	1,457	1,799
14. Switzerland	2,009	2,084	1,758	1,764	1,467	1,453
15. United Kingdom	1,748	1,806	2,301	2,230	2,315	2,212
Total	25,441	23,792	24,249	24,749	24,508	26,385

Note: Only countries that participated in all 6 ESS rounds are included in the analysis.

Table 3

Model fit indices of the multiple-group confirmatory factor analyses across 15 countries for each higher-order value and in each ESS round (configural invariance model)

	χ^2	CFI	RMSEA	SRMR
Conservation (*df* = 90)				
1st Round	433.71	.987	.046 [.042 - .050]	.017
2nd Round	356.00	.989	.042 [.037 - .046]	.016
3rd Round	323.81	.990	.038 [.034 - .043]	.015
4th Round	397.34	.988	.044 [.039 - .048]	.017
5th Round	428.85	.987	.046 [.042 - .050]	.018
6th Round	369.16	.989	.040 [.036 – 045]	.016
Self-enhancement (*df* = 15)				
1st Round	42.09	.999	.032 [.021 - .043]	.007
2nd Round	53.28	.998	.039 [.028 - .050]	.007
3rd Round	67.57	.997	.044 [.034 - .056]	.009
4th Round	52.55	.998	.037 [.027 - .049]	.008
5th Round	78.87	.997	.049 [.038 - .060]	.009
6th Round	104.64	.996	.056 [.046 - .066]	.010
Self-transcendence (*df* = 45)				
1st Round	185.36	.993	.041 [.035 - .048]	.013
2nd Round	214.82	.991	.047 [.041 - .053]	.014
3rd Round	307.53	.987	.057 [.051 - .064]	.017
4th Round	255.06	.990	.051 [045 - .057]	.015
5th Round	319.43	.988	.058 [.052 - .065]	.015
6th Round	330.80	.986	.058 [.052 - .064]	.016
Openness to change with hedonism (*df* = 90)				
1st Round	1630.7	.958	.097 [.093 - .101]	.033
2nd Round	1524.1	.957	.097 [.092 - .101]	.033
3rd Round	1665.3	.955	.099 [.095 - .104]	.034
4th Round	1415.5	.963	.090 [.086 - .095]	.031
5th Round	1264.8	.964	.085 [.081 - .090]	.030
6th Round	1394.4	.963	.087 [.083 - .091]	.030
Openness to change without hedonism (*df* = 15)				
1st Round	109.22	.994	.059 [.049 - .069]	.010
2nd Round	66.83	.997	.045 [.034 - .056]	.008
3rd Round	59.18	.997	.041 [.030 - .052]	.008
4th Round	78.74	.996	.049 [.038 - .059]	.009
5th Round	61.95	.997	.042 [.031 – 053]	.008
6th Round	51.32	.998	.036 [.025 - .047]	.007

Table 4

Global fit indices for the approximate measurement invariance tests across 15 countries for each higher-order value and in each ESS round (with a prior variance of 0.05)

	ppp	95% Credibility Interval
Self-enhancement		
1st Round of ESS	.178	-29.290 – 82.959
2nd Round of ESS	.089	-17.906 – 94.001
3rd Round of ESS	.035	-4.388 – 108.102
4rd Round of ESS	.096	-19.130 – 92.985
5th Round of ESS	.014	6.500 – 118.831
6th Round of ESS	.001	33.790 – 145.481
Self-transcendence		
1st Round of ESS	.000	72.03 – 204.32
2nd Round of ESS	.000	100.93 – 231.49
3rd Round of ESS	.000	192.84 – 323.78
4rd Round of ESS	.000	141.69 – 272.43
5th Round of ESS	.000	206.22 – 337.59
6th Round of ESS	.000	218.25 – 349.08
Conservation		
1st Round of ESS	.000	268.70 – 419.12
2nd Round of ESS	.000	188.11 – 339.10
3rd Round of ESS	.000	159.55 – 310.56
4rd Round of ESS	.000	232.02 – 383.11
5th Round of ESS	.000	263.83 – 413.27
6th Round of ESS	.000	205.24 – 356.04
Openness with hedonism		
1st Round of ESS	.000	1456.25 – 1608.96
2nd Round of ESS	.000	1351.61 – 1503.74
3rd Round of ESS	.000	1484.23 – 1637.41
4rd Round of ESS	.000	1245.37 – 1397.43
5th Round of ESS	.000	1092.85 – 1245.05
6th Round of ESS	.000	1224.73 – 1374.72
Openness without hedonism		
1st Round of ESS	.001	37.312 – 149.885
2nd Round of ESS	.035	-4.253 – 107.713
3rd Round of ESS	.059	-11.201 – 100.421
4rd Round of ESS	.014	7.054 – 119.765
5th Round of ESS	.051	-9.382 – 103.396
6th Round of ESS	.102	-19.736 – 92.334

Note: ppp = posterior predictive *p*-value.

Table 5

Global fit indices for the approximate measurement invariance tests across a subset of countries for self-transcendence and conservation in each ESS round (with a prior variance of 0.05)

	ppp	95% Credibility Interval
Self-transcendence in 12 countries: Belgium, Finland, Germany, Ireland, Netherlands, Norway, Poland, Portugal, Spain, Sweden, Switzerland, United Kingdom		
1st Round of ESS	.508	-50.05 – 49.46
2nd Round of ESS	.419	-45.17 – 55.39
3rd Round of ESS	.326	-38.57 – 61.99
4rd Round of ESS	.419	-45.13 – 55.07
5th Round of ESS	.273	-34.72 – 65.34
6th Round of ESS	.505	-48.32 – 47.27
Conservation in 10 countries: Belgium, Finland, Germany, Ireland, Netherlands, Poland, Portugal, Switzerland, United Kingdom, Slovenia		
1st Round of ESS	.173	-23.70 – 67.94
2nd Round of ESS	.131	-19.20 – 72.09
3rd Round of ESS	.135	-20.07 – 71.09
4rd Round of ESS	.097	-15.76 – 75.66
5th Round of ESS	.176	-23.61 – 66.82
6th Round of ESS	.067	-10.68 – 80.66

Note. The following items were dropped from the analysis because they were noninvariant in many countries and rounds: The importance of the environment item for the self-transcendence value; and the importance of following rules and being modest for the conservation value (a detailed technical report is available from the first author upon request). Poland had to be dropped from the analysis of self-transcendence in the 6th ESS round to reach convergence.