The comparability of measurements of attitudes toward immigration in the European Social Survey: exact versus approximate measurement equivalence

Davidov, Eldad ; Cieciuch, Jan ; Meuleman, Bart ; Schmidt, Peter ; Algesheimer, René ; Hausherr, Mirjam

Abstract: International survey datasets are analyzed with increasing frequency to investigate and compare attitudes toward immigration and to examine the contextual factors that shape these attitudes. However, international comparisons of abstract, psychological constructs require the measurements to be equivalent–i.e., they should measure the same concept on the same measurement scale. Traditional approaches to assessing measurement equivalence quite often lead to the conclusion that measurements are cross-nationally incomparable but have been criticized for being overly strict. In the current study, we present an alternative Bayesian approach that assesses whether measurements are approximately (rather than exactly) equivalent. This approach allows small variations in measurement parameters across groups. Taking a multiple group confirmatory factor analysis framework as a starting point, this study applies approximate and exact equivalence tests to the anti-immigration attitudes scale that was implemented in the European Social Survey (ESS). Measurement equivalence is tested across the full set of 271,220 individuals in 35 ESS countries over six rounds. The results of the exact and the approximate approaches are quite different. Approximate scalar measurement equivalence is established in all ESS rounds, thus allowing researchers to meaningfully compare these mean scores and their relationships with other theoretical constructs of interest. The exact approach, however, eventually proves to be overly strict and leads to the conclusion that measurements are incomparable for a large number of countries and time points.

DOI: https://doi.org/10.1093/poq/nfv008

The Comparability of Measurements of Attitudes toward Immigration in the European Social Survey: Exact Versus Approximate Measurement Equivalence

Eldad Davidov[1], Jan Cieciuch[1 2], Bart Meuleman[3], Peter Schmidt[4], René Algesheimer[1], Mirjam Hausherr[1]

[1]University of Zurich, Switzerland [2]Cardinal Stefan Wyszynski University in Warsaw, Poland

[3]University of Leuven, Belgium [4]University of Giessen, Germany

This is a pre-copy-editing, author-produced PDF of an article accepted for publication in the **Public Opinion Quarterly** *(79,S1 (2015), pp. 244–266) following peer review. It was first published online in this journal on May 06, 2015. The definitive publisher-authenticated version is available online at:*

http://poq.oxfordjournals.org/content/79/S1/244.abstract

or under

doi:10.1093/poq/nfv008

Acknowledgments:

The authors would like to thank Lisa Trierweiler, Anne Lee and Richard Bowles for the English proof of the manuscript.

Funding disclosure:

The work of the first, second, and fifth authors was supported by the University Research Priority Program 'Social Networks', University of Zurich. The work of the second author was partially supported by the Polish National Science Centre [Grant 2011/01/D/HS6/04077].

Conflict of interest

There is no conflict of interest.

*Address correspondence to Eldad Davidov, University of Zurich, Andreasstrasse 15, 8050 Zurich, +41 44 635 23 22, davidov@soziologie.uzh.ch.

The Comparability of Attitudes toward Immigration in the European Social Survey: Exact Versus Approximate Measurement Equivalence

Abstract

International survey datasets are analyzed with increasing frequency to investigate and compare attitudes toward immigration and to examine the contextual factors that shape these attitudes. However, international comparisons of abstract, psychological constructs require the measurements to be equivalent–i.e., they should measure the same concept on the same measurement scale. Traditional approaches to assessing measurement equivalence quite often lead to the conclusion that measurements are cross-nationally incomparable but have been criticized for being overly strict. In the current study, we present an alternative Bayesian approach that assesses whether measurements are *approximately* (rather than exactly) equivalent. This approach allows small variations in measurement parameters across groups.

Taking a multiple group confirmatory factor analysis framework as a starting point, this study applies approximate and exact equivalence tests to the anti-immigration attitudes scale that was implemented in the European Social Survey (ESS). Measurement equivalence is tested across the full set of 271,220 individuals in 35 ESS countries over six rounds. The results of the exact and the approximate approaches are quite different. Approximate scalar measurement equivalence is established in all ESS rounds, thus allowing researchers to meaningfully compare these mean scores and their relationships with other theoretical constructs of interest. The exact approach, however, eventually proves to be overly strict and

leads to the conclusion that measurements are incomparable for a large number of countries and time points.

Keywords: European Social Survey; approximate vs. exact measurement equivalence; attitudes toward immigration; cross-national research

Introduction

Intergroup relationships and attitudes have been the focus of scholarly attention since the early days of social science disciplines such as sociology and social psychology (e.g., Sumner 1960). However, due to substantially increasing international migration movements over the last several decades (Hooghe et al. 2008), this topic has moved notably to the front of the research agenda. The 'age of migration' (Castles and Miller 2003) and the resulting ethnic diversity–Vertovec (2007) even speaks of 'super-diversity'–have fundamentally changed the composition and outlook of the populations of Western countries. The electoral successes of anti-immigration parties in Europe (see, e.g., Anderson 1996; Lubbers, Gijsberts, and Scheepers 2002) provide evidence that the arrival of newcomers has created upheaval among substantial numbers of majority-group citizens. Perceptions that immigration has negative economic and cultural repercussions are widespread and have caused sizeable portions of Western populations to favor more restrictive immigration policies (Cornelius and Rosenblum 2005).

Numerous empirical studies have investigated the genesis of ethnic prejudice, ethnocentrism, and anti-immigration attitudes (for a historical overview, see Duckitt 1992). Ample evidence has been presented that negative attitudes toward immigration and the derogation of ethnic minority groups are systematically related to individual characteristics, such as educational level (Coenders and Scheepers 2003; Hainmueller and Hiscox 2007), individual economic interests (Citrin et al. 1997; Fetzer 2000), religiosity (McFarland 1989; Billiet 1995), human values (Sagiv and Schwartz 1995; Davidov et al. 2008), authoritarianism (Heyder and Schmidt 2003), and voting for extreme right-wing parties (Semyonov, Raijman, and Gorodzeisky 2006). More recently, scholars have also shown interest in the contextual

determinants of anti-immigration attitudes (e.g., Quillian 1995; Semyonov, Raijman, and Gorodzeisky 2006; Schneider 2008; Meuleman, Davidov, and Billiet 2009). Making use of increasingly available cross-national data sources, such as the European Social Survey (ESS) (see Jowell et al. 2007), the International Social Survey Program (ISSP), and the European Values Study (EVS), numerous papers that investigate the relationship between economic conditions, size of the immigrant population and anti-immigration feelings among the population have been published (for a review, see Ceobanu and Escandell 2010).

This 'cross-national turn' in the field of anti-immigration attitude studies has important merits, as it advances knowledge about the validity of theories in different societies and provides insights into contextual effects. At the same time, however, cross-national comparative research results in important methodological challenges (Harkness et al. 2003). Among many other methodological issues, people in different countries–with different cultural and linguistic backgrounds–may understand survey questions in diverse ways or respond in systematically different ways to the same questions. This might lead to incomparable scores and biased conclusions. Therefore, the assumption of cross-cultural measurement equivalence must be tested before cross-national comparisons are made (Meredith 1993; Vandenberg and Lance 2000; Vandenberg 2002; Harkness et al. 2010; Millsap 2011; Davidov et al. 2014).[1] In the current paper, the concept of measurement

[1] Measurement equivalence is a requirement not only in cross-national research but also applies to all possible comparisons of groups, irrespective of the characteristic that is used to delineate the groups (e.g., gender, age, educational level, religious denomination, or even cultural characteristics). Because of the diversity in economic, cultural, and linguistic backgrounds, however, cross-national designs are especially vulnerable to a lack of equivalence.

equivalence refers to the question "whether or not, under different conditions of observing and studying phenomena, measurement operations yield measures of the same attribute" (Horn and McArdle 1992: 117). Thus, measurement equivalence is a psychometric property of concrete measurements. Measurements are considered to be equivalent (i.e., eliciting equivalent responses) when they operationalize the same construct in the same manner across different groups, such as countries, regions, or cultural groups (and also conditions of data collection, time points, educational groups, etc.). When measurements are not equivalent, the risk exists that the observed similarities or differences between groups reflect measurement artifacts rather than true substantive differences. Horn and McArdle (1992) metaphorically described such a case as a comparison between apples and oranges. The presence of such measurement non-equivalence can substantially affect conclusions (see Davidov et al. 2014 for examples). Measurement equivalence is a necessary condition for applying multilevel models for cross-national data–a technique that has been frequently used in comparative anti-immigration research that utilizes survey data for the analysis (Cheung, Leung, and Au 2006). However, measurement equivalence has seldom been tested in such studies.

Various preventive measures have been developed to avoid measurement non-equivalence, and these should be applied during the phases of questionnaire development and data collection (Johnson 1998; van de Vijver 1998; Harkness et al. 2003). Among other things, accurately translated questionnaires, comparable sampling designs, and similar data collection modes should be used. However, even the most rigorous application of these standards cannot guarantee measurement equivalence. Therefore, researchers should evaluate whether the constructs that they use have been measured equivalently. Traditionally, measurement equivalence is assessed by testing whether certain parameters of a measurement model (e.g.,

factor loadings) are identical across groups. However, this approach–termed the *exact approach* in the remainder of this article–has been criticized for being overly strict. After all, cross-group differences in measurement parameters are not harmful unless they are sufficiently large to influence substantive conclusions (Meuleman 2012; Oberski 2014). The strict requirement of exact equivalence might, therefore, hastily lead to the conclusion that measurements are not comparable. To address this problem, the current study presents a Bayesian approach that tests whether measurements are *approximately* equivalent (Muthén and Asparouhov 2013; van de Schoot et al. 2013) rather than requiring measurement parameters to be exactly equivalent across countries. This alternative approach allows survey researchers to establish whether the measurement of their constructs is *sufficiently similar* across countries to allow a meaningful cross-country comparison. In the current paper, we apply the exact approach to testing for measurement equivalence and compare the results to those produced by the Bayesian procedure of approximate measurement equivalence. We focus on the most often used analytical tool to test for measurement equivalence, i.e., multiple group confirmatory factor analysis. We test the equivalence of a scale that has been used quite frequently in applied research, specifically, the ESS scale, which measures attitudes toward immigration policies. The main research questions are as follows: (1) whether the ESS measurements of anti-immigration attitudes are cross-nationally comparable and (2) whether the Bayesian approach, which assesses approximate equivalence, produces conclusions that are similar to those of the exact approach. To the best of our knowledge, this is the first study in which the approximate measurement equivalence approach is applied to large-scale survey data and compared with more traditional approaches to testing for equivalence. We begin by providing a short overview of the exact approach versus the approximate approach to test for measurement equivalence across samples. Next, we describe the data that we use and the

items that measure attitudes toward immigration. In the subsequent section, we present the results of the tests of measurement equivalence using the exact approach and the approximate approach with Bayesian estimation. The country mean scores that are computed using each of these methods are then compared with each other and with sum scores (which are the most commonly used method in substantive research to compare scores). Finally, we discuss the pros and cons of the classical exact approach versus the new approach of approximate measurement equivalence for survey research and for cross-national research in general.

An Exact Approach to Measurement Equivalence: Multiple Group Confirmatory Factor Analysis (MGCFA)

The exact approach to measurement equivalence tests whether the relationships between indicators and constructs are identical across groups. Over the last several decades, various analytical tools, such as multiple group confirmatory factor analysis (MGCFA: Jöreskog 1971; Bollen 1989; Steenkamp and Baumgartner 1998), item response theory (IRT: Raju et al. 2002), and latent class analysis (LCA: Kankaraš et al. 2011), have been proposed. Of these methods, MGCFA has likely been the most commonly used. For example, MGCFA has been used to test the cross-country equivalence of human values (Davidov et al. 2008a,b), political attitudes (Judd, Krosnick, and Milburn 1981), attitudes toward democracy and welfare policies (Ariely and Davidov 2010, 2012), social and political trust (Allum, Read, and Sturgis 2011; Delhey, Newton, and Welzel 2011; van der Veld and Saris 2011; Freitag and Bauer 2013), and national identity (Davidov 2009), to name only a few substantive applications.

The MGCFA framework for continuous data distinguishes between various hierarchically ordered levels of equivalence, each being defined by the parameters that are constrained

across groups (Steenkamp and Baumgartner 1998; Davidov, Schmidt, and Schwartz 2008).[2] Below, we discuss the three levels that are most relevant for applied researchers, namely, configural, metric, and scalar equivalence.[3] The first and lowest level of measurement equivalence is termed configural equivalence (Horn and McArdle 1992; Meredith 1993; Vandenberg and Lance 2000). Configural equivalence requires that each construct is measured by the same items. However, it remains uncertain whether the construct is measured on the same scale (Horn and McArdle 1992; Steenkamp and Baumgartner 1998; Vandenberg and Lance 2000). Metric equivalence is assessed by testing whether factor loadings are equal across the groups to be compared (Vandenberg and Lance 2000). If metric equivalence is established, a one-unit increase in the latent construct has the same meaning across all groups. Consequently, covariances and unstandardized regression coefficients may be meaningfully compared across samples (Steenkamp and Baumgartner 1998). A third and higher level of measurement equivalence is termed scalar equivalence (Vandenberg and Lance 2000). Scalar equivalence is tested by constraining the factor loadings and indicator intercepts to be equal

[2] Because the Bayesian approximate approach to equivalence can (for the moment at least) only be implemented for continuous data, we focus on the MGCFA model for continuous data in this contribution. A detailed account of equivalence testing with MGCFA for ordinal data can be found in Millsap and Yun-Tein (2004). The most important difference between the two models is that the latter includes an additional set of parameters, namely, thresholds that link the indicators to what are termed latent response variables. The presence of these additional parameters has consequences for the levels of measurement equivalence that are distinguished and their operationalization.

[3] In addition to these three, various other levels of measurement equivalence can be defined. Steenkamp and Baumgartner (1998), for example, also distinguished levels that imply the equality of residual variances and variances and covariances of the latent factors. Because these levels have fewer practical implications, we do not discuss them in detail here.

across groups (Vandenberg and Lance 2000). The establishment of scalar equivalence implies that respondents with the same value on the latent construct have the same expected response, irrespective of the group that they belong to. As a consequence, latent means can also be compared across groups because the same construct is measured in the same manner.

In practice, it can be quite difficult to reach measurement equivalence, especially the higher levels (Asparouhov and Muthén 2014). Variations in how respondents react to specific question wordings or survey questions in general (i.e., social desirability or 'yes-saying' tendency) can be affected by cultural or national backgrounds. Therefore, such variations might distort responses to the extent that scalar equivalence is not supported, particularly in cross-national data but also within countries, especially when there are language or cultural differences among groups (see, for example, Davidov et al. 2008; Meuleman and Billiet 2012). In certain situations, the concept of partial equivalence can offer a solution. Byrne, Shavelson and Muthén (1989) argued that not all indicators of a concept must perform equivalently across all groups. Partial equivalence implies that at least two indicators should have equal measurement parameters (i.e., loadings for partial metric equivalence and loadings plus intercepts for partial scalar equivalence). When at least two such comparable 'anchor items' are present, differential item functioning in other items can be corrected for and meaningful comparisons across groups remain possible. It is important to note, however, that this notion of partial equivalence remains within the framework of the exact approach to measurement equivalence. For at least two indicators, parameters are required to be identical across groups (while the parameters for other indicators can vary to a great extent). This is a crucial difference from the approximate approach that is explained in the next section. In this approach, the measurements for all indicators are allowed to vary minimally.

In literature concerning MGCFA, there are two common approaches to evaluate whether measurement parameters are identical across groups (the two approaches do not exclude each other and can be applied simultaneously). The first approach relies on various global fit indices (Chen 2007). The second approach focuses on detecting local misspecifications (Saris, Satorra, and van der Veld 2009).

In the first approach, various global fit indices are used to assess the correctness of the model. In addition to the chi-square test (which has been criticized because of its sensitivity to sample size), the following three alternative fit indices are quite frequently mentioned in the relevant literature: the root mean square error of approximation (RMSEA), the comparative fit index (CFI), and the standardized root mean square residual (SRMR). To assess whether a given level of measurement equivalence has been established, global fit measurements are compared between more and less constrained models. If the change in model fit is smaller than the criteria that are proposed in the literature, measurement equivalence for that level is established. According to a simulation study by Chen (2007), if the sample size is larger than 300, metric non-equivalence is indicated by a change in CFI larger than .01 when supplemented by a change in the RMSEA larger than .015 or a change in SRMR larger than .03 compared with the configural equivalence model. With regard to scalar equivalence, non-equivalence is evidenced by a change in CFI larger than .01 when supplemented by a change in RMSEA larger than .015 or a change in SRMR larger than .01 compared with the metric equivalence model.

In the second approach, the evaluation of the model correctness is based on the determination of whether any local misspecifications are present in the model rather than on an assessment of global fit. A correct model should not contain any relevant misspecifications. In the context of equivalence testing, possible misspecifications include factor loadings or item intercepts that are incorrectly set equal across countries. According to Saris, Satorra, and van der Veld (2009), it is possible for the global fit criteria to indicate the satisfactory fit of a model, although the model contains serious misspecifications and, consequently, should be rejected. It is also possible that although the global fit measurements suggest that a model should be rejected, it may not contain any relevant misspecifications and, accordingly, should be accepted (Saris et al. 2009). The second case is particularly likely to occur with models that are rather complex or that contain many groups.

Saris, Satorra, and van der Veld's (2009) recommendation consists of the following two elements: 1) to rely on modification indexes (MI), which provide information on the minimal decrease in the chi-square of a model when a given constraint is released, and on the expected parameter change (EPC) that is provided in the output and 2) to take into account the power of the modification index test. Neither the EPC nor the MI test is free of problems. The EPC estimation is problematic because sampling fluctuations may influence it. In addition, the value of the EPC depends on other misspecifications in the model. To resolve this problem, Saris et al. (2009) introduced the standard error of the EPC and the power of the MI test. According to Saris et al. (1987), both the standard error of the EPC and the power can be estimated based on the MI and EPC. Saris et al. (2009) suggested that the correct model should not contain any relevant misspecifications, whereas every serious misspecification is an indicator of the necessity to either reject or modify the model. An important feature of this

approach is that the researcher defines the threshold at which misspecification requires detection. Saris, Satorra, and van der Veld (2009) suggested treating deviations larger than .4 for cross-loadings and deviations larger than .1 for differences in factor loadings or intercepts across groups as misspecified (for further details, we refer readers to the Saris et al. 2009 study).

Problems with the Exact Approach

As previously indicated, in many cases, it is not possible to establish full or even partial cross-cultural equivalence with survey research data (Davidov et al. 2008b; Meuleman and Billiet 2012; Asparouhov and Muthén 2014; for a review, see Davidov et al. 2014). This implies that measurement parameters, such as loadings or intercepts, are not identical across groups. This finding may preclude any meaningful comparisons across groups under study because researchers cannot guarantee that the comparisons are valid. Van de Schoot et al. (2013) metaphorically described this problem as "traveling between Scylla and Charybdis", which refers to having to choose between two evils. Scylla represents a model with imposed equality constraints that fits the data poorly, whereas Charybdis represents a model that fits the data well but contains no equality constraints. Both "monsters" are threatening, and the danger lies in the fact that the researcher cannot know whether the differences between groups (such as cultures, countries, geographical areas, or language groups within a country) are due to real differences or due to methodological artifacts (i.e., measurement inequivalence). Van de Schoot et al. (2013) proposed following a third option for "traveling between Scylla and Charybdis", specifically, applying the approximate Bayesian measurement equivalence approach.

The Bayesian Approach for Establishing Approximate Measurement Equivalence across Groups

The procedure that constrains parameters (e.g., factor loadings and intercepts) to be exactly equal to establish measurement equivalence is quite demanding. It can legitimately be questioned whether it is necessary for measurement parameters to be completely identical across groups to allow meaningful comparisons. It is possible that 'nearly equal' is sufficient to guarantee that comparisons are unbiased, assuming that 'nearly can be operationalized. Such a consideration underlies the Bayesian approach to measurement equivalence, recently implemented by Muthén and Asparouhov (2012, 2013) in the Mplus software package (Muthén and Muthén 1998-2012). According to this approach, approximate rather than exact measurement equivalence can be tested. Approximate measurement equivalence permits small differences between parameters that would otherwise be constrained to be equal in the traditional exact approach for testing measurement equivalence. The parameters that are specified in a Bayesian approach are considered to be variables, and their distribution is described by prior probability distribution (PPD). A researcher can introduce their knowledge or assumptions about the PPDs into the analysis and can define them (Muthén and Asparouhov 2013; Davidov et al. 2014). More specifically, when testing for measurement equivalence, a researcher may expect differences between factor loadings or intercepts across groups to be zero but may wish to allow their differences to vary slightly across groups. Simulations have suggested that small variations may be allowed without risking invalid conclusions in comparative research (van de Schoot et al. 2013). The evaluation of the model

should detect whether actual deviations from equality across groups exceed these limits suggested by simulation studies.[4]

The fit of the Bayesian model can detect whether actual deviations are larger than those that the researcher allows in the prior distribution. A Posterior Predictive p-value (PPP) of a model can be obtained based on the usual likelihood-ratio chi-square test of an H0 model against an unrestricted H1 model. A low PPP indicates a poor fit (Muthén and Asparouhov 2012). If the prior variance is small relative to the magnitude of non-invariance, the PPP will be lower than if the prior variance corresponds more closely to the magnitude of non-invariance. The model fit can also be evaluated based on the credibility interval (CI) for the difference between the observed and the replicated chi-square values. According to Muthén and Asparouhov (2012) and van de Schoot et al. (2013), the Bayesian model fits to the data when the PPP is larger than zero and the CI contains zero. Additionally, Mplus lists all parameters that significantly differ from the priors. This feature is similar to modification indices in the exact measurement invariance approach. Although the model is assessed based on PPP and CI, these values provide global model fit criteria that are similar to the criteria in the exact approach (Chen 2007).

The Current Study

Several studies have demonstrated that it is rather difficult to reach scalar and, at times, even metric levels of measurement equivalence when tested on large-scale survey data that

[4] To avoid a situation in which researchers 'trim' their model to find the optimal priors that ensure equivalence, simulation studies provide guidelines as to how large these priors may be (van de Schoot et al. 2013). We rely on these studies in the empirical portion of the current paper.

includes many countries or other cultural groups (Asparouhov and Muthén 2014; Davidov et al. 2014). The Bayesian approximate equivalence approach is promising, as it may suggest that groups are comparable and that their scores may be meaningfully compared even when traditional exact approaches suggest that this is not possible. However, Bayesian analysis for assessing measurement equivalence is a newly implemented approach (Muthén and Asparouhov 2013); therefore, knowledge is quite limited concerning how the results of Bayesian approximate measurement equivalence compare with the results of traditional exact measurement equivalence approaches. The current study is the first to empirically compare the findings of measurement equivalence analyses using the exact approach and the Bayesian approach of approximate measurement equivalence. This study investigates whether, in practice, Bayesian analysis may provide findings that allow substantive survey researchers to meaningfully compare scores across countries even when an assessment of exact equivalence would not allow such a comparison.

For the analysis, we employ a large dataset from six rounds of the European Social Survey (ESS), which measures attitudes toward immigration policies. The ESS is a biennial cross-national European survey that is administered to representative samples from approximately 30 countries. Since its inception in 2002/2003, its core module has included questions that measure attitudes toward immigrants and immigration policies. These questions have been repeated in each round and used extensively in cross-national research in over 60 publications to date, including some published in highly ranked journals. Thus, these questions largely contribute to immigration research and policy debates (Heath et al. 2014). In such a large-scale survey, it is crucial to determine whether scores based on these measurements may be meaningfully compared across countries. We assess their comparability using the Bayesian

approximate invariance approach and compare the findings with those using the exact approach in the next section.

METHODS

Data and Measurements

A total of 35 countries and six rounds of the ESS (2002/3, 2004/5, 2006/7, 2008/9, 2010/11, and 2012/13) are included in the study. Not all countries participated in all rounds. Some countries joined early in 2002/3 and did not participate in later rounds. Other countries did not participate in the ESS at the beginning but joined later. After excluding respondents whose country of birth was not the same as their residence, the total sample size is 271,220 respondents. Table 1 summarizes the number of participants in each round who are included in the analysis. The data were retrieved from the ESS website (www.europeansocialsurvey.org). Further information on data collection procedures, the full questionnaire, response rates, and methodological documentation is available on the ESS website.

Table 1 here

Three items in the ESS measure attitudes toward immigration policies. They are formulated as follows: (1) “To what extent do you think [country] should allow people of the same race or ethnic group from most [country] people to come and live here?” (2) “To what extent do you think [country] should allow people of a different race or ethnic group from most [country, adjective form] people to come and live here?” and (3) “To what extent do you think [country] should allow people from the poorer countries outside Europe to come and live

here?" The respondents recorded their responses to these three questions on 4-point scales ranging from 1 "allow none" to 4 "allow many".

Plan of Analysis

1. Testing for exact (full or partial) equivalence

First, we ran six MGCFA analyses using the full information maximum likelihood (FIML) procedure (Schafer and Graham 2002), one for each round, with all of the countries included in the particular round. Each analysis contained three separate assessments of configural, metric, and scalar equivalence, with the corresponding constraints for the metric and scalar levels of measurement equivalence. To identify the model, we used the second approach that was proposed by Little, Slegers, and Card (2006), termed the marker-variable method, and constrained the loading of one of the items to 1 and the intercept of this item to 0 for all countries. If the loading and/or intercept of this item varied considerably across countries, we used a different reference item for identification. If full measurement equivalence was not established, we attempted to assess partial measurement equivalence. We used the program Jrule (Oberski 2009; Saris et al. 2009) to detect local misspecifications of parameters whose equality constraint should be released according to the program. To establish partial scalar equivalence, only one item could be released because partial scalar equivalence requires that the parameters of at least two items are constrained to be equal across all groups. However, as shown in the next section, the results of the analyses using Jrule indicated misspecifications for two or even three items in several countries. This result indicated that in these countries, even partial scalar equivalence was not established.

2. Testing for approximate scalar equivalence

The assessment of approximate measurement equivalence using Bayesian analysis requires imposing priors on specific parameters. When testing for approximate measurement equivalence, the *average* difference between loadings and intercepts across countries is assumed to be zero, as in MGCFA when testing for exact measurement equivalence. However, there is one difference, i.e., approximate measurement equivalence permits small variations between parameters that would be constrained to be exactly equal in the traditional exact approach for testing for measurement equivalence. Using simulation studies, van de Schoot et al. (2013) demonstrated that variance as large as 0.05 imposed on the difference between the loadings or the intercepts does not lead to biased conclusions when approximate equivalence is assessed. We followed their recommendations and imposed the following priors on the difference parameters of the loadings and intercepts: mean difference = 0 and variance of the difference = .05. We used similar constraints to identify the model as in the MGCFA. Specifically, we constrained the loading of one item to (exactly) 1 in all groups and the intercept of this item to (exactly) 0 in all groups. If the loading and/or intercept of this item varied considerably across countries, we chose a different reference item to use for identification. The latent means and variances were freely estimated in all countries.

3. Comparison of the obtained results

We compared the country means that were obtained from the exact and Bayesian analyses with each other as well as with those based on the raw sum scores. We estimated the correlation between the country rankings based on each of the three procedures in each ESS round.

RESULTS

We first ran MGCFA to assess exact measurement equivalence across countries in each round. Figure 1 displays the model that we tested. This model includes a latent variable that measures attitudes toward immigration policies with three items. Table 2 summarizes the global fit measurements for sequentially more constrained models for this latent variable in each ESS round.

Figure 1 here

Equivalence in the exact approach. As Table 2 illustrates, the changes in CFI for the metric equivalence level (compared with the configural level) were less than 0.01, indicating that they are acceptable. However, changes for SRMR and RMSEA exceeded the recommended cut-off criteria (namely, 0.015 and 0.03, respectively; Chen 2007). The results revealed (in the analysis performed by Jrule) that the factor loading of one item–measuring whether respondents wished their country to allow entry to many or few immigrants of the same race or ethnic group as the majority–considerably differed across countries in all rounds repeatedly. Therefore, we released the constraint on this factor loading and tested for partial metric equivalence. Following this modification, two of the fit indices (CFI and SRMR) indicated an acceptable fit between the model and the data in all rounds, which was satisfactory for not rejecting the partial metric equivalence model (Meuleman and Billiet 2012). Thus, according to these measurements, the data supported partial metric equivalence for all rounds. This finding implies that the meaning of the construct measuring attitudes toward immigration policies is likely similar across countries. This finding is, however, not sufficient to allow comparisons of this attitude's means across countries. Mean comparisons require a higher level of equivalence, specifically, partial or full scalar equivalence.

Table 2 here

We next tested for partial scalar equivalence. We constrained the factor loadings and intercepts of two items to be equal across all countries in each round while allowing both the factor loading and the intercept of the item measuring whether respondents wished their country to allow many or few immigrants of the same race or ethnic group to be freely estimated. Table 2 summarizes the global fit measurements for this test in each ESS round. As Table 2 illustrates, the changes in CFI and SRMR for the partial scalar equivalence model (compared with the partial metric equivalence model) were relatively acceptable. However, those for RMSEA were acceptable only for the data in the first round. In all other rounds, the changes in RMSEA exceeded the cut-off criteria that Chen (2007) recommended. In addition, the intercept of one or two more items varied considerably across several countries. Jrule helped us to identify those items. Therefore, we concluded that the scale did not meet the requirements of partial scalar equivalence based on this criterion across the *full set* of ESS countries. However, at times, researchers are interested in comparing a subset of the countries, and partial scalar equivalence may hold for subsets of countries. Therefore, mean comparisons of attitudes toward immigration can be made across the countries in the subset. Table 3 lists the countries where partial scalar equivalence was not supported by the data in each round. For example, in the second ESS round, Estonia, Portugal, Slovenia, and Ukraine did not reach partial scalar equivalence. This finding implies that the means of attitudes toward immigration may be compared across all of the other countries in this round. It should be noted that although the global fit measurements suggest that the means may be compared across all countries in the first round, Jrule identified two countries where this was not the case, i.e., Hungary and Israel. The respondents seemed to react differently to the immigration

questions in these two countries. As a result, their scores were not comparable with those of respondents in other countries. The largest share of non-comparable countries was found in the sixth ESS round. On average, 30 percent of the ESS countries were not comparable on the attitudes toward immigration score. This result is quite disappointing because it may preclude meaningful mean comparisons across a large proportion of the ESS countries.[5] Accordingly, it may be questioned whether the strict assumption of exact measurement equivalence is *necessary* to conduct meaningful comparisons. Next, we loosen this assumption by turning to a test of approximate measurement equivalence.

Table 3 here

*Equivalence in the approximate approach***.** Our second research question was whether Bayesian analyses, which assess approximate equivalence, establish higher levels of equivalence. Table 4 presents the model fit coefficients for the approximate Bayesian analyses.

[5] Lubke and Muthén (2004) criticized the analysis of Likert data under the assumption of normality. They proposed that in such a case, a model should be fitted for ordered categorical outcomes. Indeed, we made the assumption that the data are continuous, although ordinal categorical. This is a common assumption when the sample size is large. However, the items in our analysis have only four points (rather than the more common five points) on the scale. Therefore, we re-ran the exact approach taking into account the ordinal-categorical character of the data. The findings remained essentially the same and are provided in the Appendix. They suggest that equivalence cannot be supported across all countries in all six rounds based on the exact approach. Unfortunately, at this time, a Bayesian analysis that considers the ordinal-categorical character of the data while including thresholds in the model is unavailable.

Table 4 here

The findings revealed that approximate scalar measurement equivalence was established across *all* countries in all ESS rounds. All PPP values were higher than zero, and the 95 percent CI for the difference between the observed and the replicated chi-square values contained zero (Muthén and Asparouhov 2012, 2013). These global fit measurements are sufficient to accept the model and, thus, allow the comparison of the scores of attitudes toward immigration across all countries in each round of the ESS (van de Schoot et al. 2013), although the exact approach failed to do so.

Comparison of the obtained results. The results of the exact and approximate measurement equivalence are quite different. Approximate scalar measurement equivalence was established in each ESS round separately, whereas exact scalar measurement equivalence (across all countries) was not established in all ESS rounds. However, if the measurement is sufficiently equivalent across countries to conduct meaningful comparisons, as indicated by the approximate procedure, the latent means estimated in the exact MGCFA should be trustworthy as well, although the exact MGCFA failed to establish even partial scalar measurement equivalence (Muthén and Asparouhov 2013). To examine this, we estimated mean scores based on the exact and approximate approaches and compared them to each other and to sum scores that were computed using the raw data. As many substantive and applied survey researchers are more interested in the country rankings than the means, we next ranked the countries based on the means that were obtained in each procedure and calculated the correlations between these rankings for each ESS round. Table 5 lists the correlations between the country rankings and each method.

Table 5 here

As clearly shown in Table 5, all of the correlations were very high (> .95). In other words, the rankings of the means that were obtained in each of the three procedures were quite similar. However, they were much more similar to each other when estimated based on the latent variable scores in the approximate and exact approaches than to those estimated based on the sum scores. This implies that the use of means based on the sum scores might, at times, lead to incorrect conclusions about the country rankings of attitudes toward immigration in the ESS data. This is an encouraging result for applied researchers. Although strictly speaking, exact scalar measurement equivalence was not supported across all countries, approximate equivalence was established. This result implies that the latent variable means that were estimated based on the approximate and exact approaches were comparable after all. However, it should be noted that such an encouraging result might not necessarily be established for other scales. It is possible that both exact and approximate approaches fail to demonstrate cross-country equivalence. In that case, various strategies, such as attempting to identify subgroups of countries and indicators for which equivalence holds or attempting to explain why certain measures lack equivalence, are available (for further details, see Davidov et al. 2014).

SUMMARY AND CONCLUSIONS

In most published cross-national studies, metric and scalar measurement equivalence is implicitly assumed without testing this assumption. This may lead to biased mean comparisons and biased comparisons of covariances and regression coefficients (Vandenberg

and Lance 2000; Kuha and Moustaki 2013; Oberski 2014). However, the traditional estimation procedures in MGCFA to test for measurement equivalence and the corresponding global fit measurements–such as chi-square difference tests, CFI differences, RMSEA differences, SRMR differences, or other common criteria (e.g., those implemented in the Jrule program)–often lead, especially in the case of scalar equivalence assessments, to a rejection of the assumption of even partial scalar equivalence. This is especially the case when data from different countries or cultures are compared and frequently results in a considerable reduction in the number of countries that can be meaningfully compared on the basis of means (Byrne and van de Vijver 2010). This can be demonstrated in the current study on the comparability of the attitudes toward immigration within six rounds of the European Social Survey between 2002 and 2012. Using the traditional procedures to test for metric and scalar equivalence leads to the incorrect (and likely overly conservative) conclusion that one needs to omit 30 percent of the countries, on average, because their mean scores on the scale might not be comparable.

To solve this problem, we applied the newly proposed procedure 'approximate measurement equivalence', which allows variance around the point estimates for the factor loadings and intercepts of the indicators. To perform this procedure, we use the Bayesian estimation framework. Muthén and Asparouhov (2012) and van de Schoot et al. (2013) proposed this framework as an alternative estimation procedure to check for measurement equivalence of multiple indicators and unbiased estimation of latent means. In the six rounds of the ESS, we demonstrated that the assumption of approximate metric and scalar equivalence was tenable using this alternative, more flexible procedure. As a consequence, the latent means of attitudes toward immigration can be legitimately compared over countries in the six time

points. The exact approach eventually proved to be overly strict and led to the conclusion that such a comparison might not be possible across countries. Therefore, researchers may now use ESS data to evaluate attitudes toward immigration across the ESS countries. The findings of cross-country approximate equivalence allow the confident comparison of these latent scores across countries and their use in comparative studies.

This study has several limitations. First, it is not clear whether the ordinal nature of the outcomes might affect the results. Whereas exact measurement invariance tests can take the ordinal character of item scores into account in the estimation, the Bayesian approach does not appropriately address this problem and assumes that scores are continuous. Future research should address this problem by developing Bayesian procedures that allow testing for approximate measurement invariance while taking into account the ordinal character of the data. Second, the amount of the variance that is specified for the priors remains to be explored. Based on previous recommendations (van der Schoot et al. 2013), we set a small magnitude of .05 or lower to establish invariance. Specifying an overly small variance may result in failure to establish invariance, whereas specifying an overly large variance may lead to incorrectly establishing invariance. Therefore, further simulations are necessary to more precisely determine the magnitude of the variance that may be specified for the priors. Finally, the level of PPP that should be considered as supportive of approximate measurement invariance has not been fully determined. Muthén and Asparouhov (2013) indicated that the PPP should be higher than zero; however, more concrete recommendations are required.

In summary, an equivalence test should be conducted to assess comparability when countries or other groups are compared. The failure to guarantee equivalence may imply that

comparability is not a given. However, approximate equivalence testing may succeed in establishing equivalence when traditional (exact) approaches fail. Using the words of van de Schoot et al. (2013), there may be a third route between Scylla and Charybdis in cross-country equivalence testing. The two 'monsters' may not necessarily be dangerous, as our case has illustrated, and may produce trustworthy means, as we have demonstrated here. Of note, however, the Bayesian test of approximate invariance cannot establish approximate invariance when measurements are completely different; it does not perform 'magic'. However, it can inform researchers when measurements are *sufficiently similar* to allow meaningful substantive comparisons. Building on these findings, a systematic equivalence test that uses various methods for other scales in the ESS and in other large data-generating programs is desirable in conducting future meaningful cross-national comparisons.

Appendix

Global Fit Measurements for the Ordinal Multiple Group CFA

Level of invariance	Chi^2	df	RMSEA	WRMR	CFI
1st Round of ESS					
Full (metric and scalar)	5717.5	147	.150 (.147-.153)	17.1	.992
Partial	2633.7	84	.134 (.130-.139)	11.0	.996
2nd Round of ESS					
Full (metric and scalar)	6043.9	175	.141 (.138-.144)	17.1	.991
Partial	2323.8	100	.115 (.111-.119)	9.9	.997
3rd Round of ESS					
Full (metric and scalar)	7606.5	168	.162 (.159-.165)	19.4	.989
Partial (1)	1953.1	96	.107 (.103-.111)	9.2	.997
4rd Round of ESS					
Full (metric and scalar)	10990.2	210	.172 (.169-.175)	23.5	.987
Partial (2)	2728.6	120	.112 (.108-.116)	10.8	.997
5th Round of ESS					
Full (metric and scalar) (3)	9743.6	182	.175 (.172-.178)	22.1	.989
Partial (4)	2662.6	104	.120 (.116-.124)	10.5	.997
6th Round of ESS					
Full (metric and scalar) (5)	6263.1	161	.150 (.146-.153)	17.7	.992
Partial (5)	1912.9	92	.108 (.104-.112)	9.1	.998

Notes. CFA = confirmatory factor analysis. Partial = released equality constraints on the thresholds of the item “allow people of the same race or ethnic group to come into the country”. This modification is consistent with the modification that is required on the factor loadings and indices in the continuous case.

References

Allum, Nick, Sanna Read, and Patrick Sturgis. 2011. "Evaluating Change in Social and Political Trust in Europe using Multiple Group Confirmatory Factor Analysis with Structured Means." In *Methods for Cross-Cultural Analysis: Basic Strategies and Applications,* edited by Eldad Davidov, Peter Schmidt, and Jaak Billiet, 37–55. London: Taylor & Francis.

Anderson, Christopher J. 1996. "Economics, Politics, and Foreigners: Populist Party Support in Denmark and Norway." *Electoral Studies* 15:497–511.

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* 104:271–86.

------. 2012. "Assessment of Measurement Equivalence with Cross-National and Longitudinal Surveys in Political Science". *European Political Science* 11:363-77.

Asparouhov, Tihomir, and Bengt Muthén. 2014. "Multiple-Group Factor Analysis Alignment." *Structural Equation Modeling* 21:1-14.

Billiet, Jaak B. 1995. "Church Involvement, Individualism, and Ethnic Prejudice among Flemish Roman Catholics: New Evidence of a Moderating Effect." *Journal for the Scientific Study of Religion* 34:224–33.

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

Byrne, Barbara M., Richard. J. Shavelson, and Bengt O. Muthen. 1989. "Testing for the Equivalence of Factor Covariance and Mean Structures - the Issue of Partial Measurement Invariance." *Psychological Bulletin* 105:456–66.

Byrne, Barbara, M., and Fons J. R. van de Vijver. 2010. "Testing for Measurement and Structural Equivalence in Large-Scale Cross-Cultural Studies: Addressing the Issue of Nonequivalence." *International Journal of Testing* 10:107-32.

Castles, Stephen, and Mark J. Miller. 2003. *The Age of Migration: International Population Movements in the Modern World.* Basingstoke: Palgrave Macmillan.

Ceobanu, Alin M., and Xavier Escandell. 2010. "Comparative Analyses of Public Attitudes toward Immigrants and Immigration using Multinational Survey Data: A Review of Theories and Research." *Annual Review of Sociology* 36:309–28.

Chen, Fang F. 2007. "Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance." *Structural Equation Modeling* 14:464–504.

Cheung, Mike W-L., Kwok Leung, and Kevin Au. 2006. "Evaluating Multilevel Models in Cross-Cultural Research. An Illustration with Social Axioms." *Journal of Cross-Cultural Psychology* 37:522-41.

Citrin, Jack, Donald P. Green, Christopher Muste, and Cara Wong. 1997. "Public Opinion toward Immigration Reform: the Role of Economic Motivations." *The Journal of Politics* 59:858–81.

Coenders, Marcel, and Peer Scheepers. 2003. "The Effect of Education on Nationalism and Ethnic Exclusionism: an International Comparison." *Political Psychology* 24:313–43.

Cornelius, Wayne A., and Marc R. Rosenblum. 2005. "Immigration and Politics." *Annual Review of Political Science* 8:99–119.

Davidov, Eldad. 2009. "Measurement Equivalence of Nationalism and Constructive Patriotism in the ISSP: 34 Countries in a Comparative Perspective." *Political Analysis* 17:64–82.

Davidov, Eldad, Bart Meuleman, Jaak Billiet, and Peter Schmidt. 2008. "Values and Support

for Immigration. A Cross-Country Comparison." *European Sociological Review* 24:58–99.

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.

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.

Delhey, Jan, Kenneth Newton, and Christian Welzel. 2011. "How General is Trust in 'Most People'? Solving the Radius of Trust Problem." *American Sociological Review* 76:786–807.

Duckitt, John H. 1992. "Psychology and Prejudice: A Historical Analysis and Integrative Framework." *American Psychologist* 47:1182–93.

Fetzer, Joel S. 2000. "Economic Self-Interest or Cultural Marginality? Anti-Immigration Sentiment and Nativist Political Movements in France, Germany and the USA." *Journal of Ethnic and Migration Studies* 26:5–23.

Freitag, Markus, and Paul C. Bauer. 2013. "Testing for Measurement Equivalence in Surveys. Dimensions of Social Trust across Cultural Contexts." *Public Opinion Quarterly* 77:24–44.

Hainmueller, Jens, and Michael J. Hiscox. 2007. "Educated Preferences: Explaining Attitudes toward Immigration in Europe." *International Organization* 61:399–442.

Harkness, Janet A, Michael Braun, Brad Edwards, Timothy P. Johnson, Lars Lyberg, Peter Ph. Mohler, Beth-Ellen Pennell, and Tom W. Smith, eds. 2010. *Survey Methods in*

Multinational, Multiregional, and Multicultural Contexts. Hoboken NJ: John Wiley & Sons.

Harkness Janet A., Fons J.R. van de Vijver, and Peter P. Mohler, eds. 2003. *Cross-Cultural Survey Methods.* New York: John Wiley

Heath, Anthony, Peter Schmidt, Eva Green, Alice Ramos, Eldad Davidov, and Robert Ford. 2014. "Attitudes towards Immigration and their Antecedents. ESS7 Rotating Modules." Retrieved from http://www.europeansocialsurvey.org/methodology/questionnaire/ESS7_rotating_module_immigration.html

Heyder, Aribert and Peter Schmidt. 2003. "Authoritarianism and Ethnocentrism in East and West Germany - Does the system matter?" In *Germans or Foreigners? Attitudes Toward Ethnic Minorities in Post-Reunification Germany,* edited by R. Alba, P. Schmidt, and M. Wasmer, 97-104. New York: Palgrave, St. Martin's Press.

Hooghe, Marc, Ann Trappers, Bart Meuleman, and Tim Reeskens. 2008. "Migration to European Countries. A Structural Explanation of Patterns, 1980-2004." *International Migration Review* 42:476-504.

Horn, John L., and John J. McArdle. 1992. "A Practical and Theoretical Guide to Measurement Invariance in Aging Research." *Experimental Aging Research* 18:117–44.

Jöreskog, Karl G. 1971. "Simultaneous Factor Analysis in Several Populations." *Psychometrika* 36:409–26.

Johnson, Timothy P. 1998. "Approaches to Equivalence in Cross-Cultural and Cross-National Survey Research. In *Zuma-Nachrichten Spezial Volume 3. Cross-Cultural Survey Equivalence*, edited by J. A. Harkness, 1–40. Mannheim, Germany: Zuma.

Jowell, Roger, Caroline Roberts, Rory Fitzgerald, and Gillian Eva. 2007. *Measuring Attitudes Cross-Nationally. Lessons from the European Social Survey*. London: Sage

Judd, Charles M., Jon A. Krosnick, and Michael A. Milburn. 1981. "Political Involvement and Attitude Structure in the General Public." *American Sociological Review* 46:660–69.

Kankaraš Milos, Jeroen K. Vermunt, and Guy Moors. 2011. "Measurement Equivalence of Ordinal Items: A Comparison of Factor Analytic, Item Response Theory, and Latent Class Approaches." *Sociological Method and Research* 40:279-310

Kuha, Jouni, and Irini Moustaki. 2013. "Non-Equivalence of Measurement in Latent Variable Modelling of Multigroup Data: A Sensitivity Analysis." Unpublished manuscript retrieved from website: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2332071

Little, Todd D., David W. Slegers, and Noel A. Card. 2006. "A Non-Arbitrary Method of Identifying and Scaling Latent Variables in SEM and MACS Models." *Structural Equation Modeling* 13:59–72.

Lubbers, Marcel, Merové Gijsberts, and Peer Scheepers. 2002. "Extreme Right-Wing Voting in Western Europe." *European Journal of Political Research* 41:345–78.

Lubke, Gitta H., and Bengt O. Muthén. 2004. "Applying Multigroup Confirmatory Factor Models for Continuous Outcomes to Likert Scale Data Complicates Meaningful Group Comparisons." *Structural Equation Modeling* 11:514–34.

McFarland, Sam G. 1989. "Religious Orientations and the Targets of Discrimination." *Journal for the Scientific Study of Religion* 28:324–36.

Meredith, William. 1993. Measurement Invariance, Factor Analysis and Factorial Invariance. *Psychometrika* 58:525–43.

Meuleman, Bart. 2012. "When are Intercept Differences Substantively Relevant in Measurement Invariance Testing." In *Methods, Theories, and Empirical Applications in*

the Social Sciences: Festschrift for Peter Schmidt, edited by S. Salzborn, E. Davidov, and J. Reinecke, 97–104. Heidelberg: Springer VS

Meuleman, Bart, and Jaak Billiet. 2012. "Measuring Attitudes toward Immigration in Europe: The Cross-Cultural Validity of the ESS Immigration Scales." *ASK: Research & Methods* 21:5–29.

Meuleman, Bart, Eldad Davidov, and Jaak Billiet. 2009. "Changing Attitudes toward Immigration in Europe, 2002-2007: A Dynamic Group Conflict Theory Approach." *Social Science Research* 38:352–65.

Millsap, Roger E. 2011. *Statistical Approaches to Measurement Invariance*. New York: Taylor and Francis Group.

Millsap, Roger E., and Jenn Yun-Tein. 2004. "Assessing Factorial Invariance in Ordered-Categorical Measures." *Multivariate Behavioral Research* 39:479–515.

Muthén, Bengt O., and Tihomir Asparouhov. 2012. "Bayesian Structural Equation Modeling: A More Flexible Representation of Substantive Theory." *Psychological Methods* 17:313–335.

Muthén, Bengt O., and Tihomir Asparouhov. 2013. "BSEM Measurement Invariance Analysis." Mplus Web Notes: No. 17. www.statmodel.com. Accessed: July 30, 2014

Muthén, Linda, and Bengt O. Muthén. 1998-2012. *Mplus User's Guide. Version 7.* Los Angeles, CA: Muthén & Muthén.

Oberski, Daniel. 2009. Jrule for Mplus version 0.91 (beta). Available at https://github.com/daob/JruleMplus/wiki Accessed: July 30, 2014

Oberski, Daniel L. 2014. "Evaluating Sensitivity of Parameters of Interest to Measurement Invariance in Latent Variable Models." *Political Analysis* 22:45–60.

Quillian, Lincoln. 1995. "Prejudice as a Response to Perceived Group Threat: Population

Composition and Anti-Immigrant and Racial Prejudice in Europe." *American Sociological Review* 60:586–611.

Raju Nambury S., Larry J. Laffitte, and Barbara M. Byrne. 2002. "Measurement Equivalence: A Comparison of Methods Based on Confirmatory Factor Analysis and Item Response Theory." *Journal of Applied Psychology* 87:517-29

Sagiv, Lilach, and Shalom H. Schwartz. 1995. "Value Priorities and Readiness for Out-Group Social Contact." *Journal of Personality and Social Psychology* 69:437–48.

Saris, Willem E., Albert Satorra, and Dag Sörbom. 1987. "The Detection and Correction of Specification Errors in Structural Equation Models." *Sociological Methodology* 17:105–29.

Saris, Willem E., Albert Satorra, and William. M. van der Veld. 2009. "Testing Structural Equation Models or Detection of Misspecifications?" *Structural Equation Modeling,* 16:561–82.

Schafer, Joseph L., and John W. Graham. 2002. "Missing Data: Our View of the State of the Art." *Psychological Methods* 7:147–77.

Schneider, Silke L. 2008. "Anti-Immigrant Attitudes in Europe: Outgroup Size and Perceived Ethnic Threat." *European Sociological Review* 24:53–67.

Semyonov, Moshe, Rebeca Raijman, and Anastasia Gorodzeisky. 2006. "The Rise of Anti-Foreigner Sentiment in European Societies, 1988–2000." *American Sociological Review* 71:426–49.

Steenkamp, Jan-Benedict E. M., and Hans Baumgartner. 1998. "Assessing Measurement Invariance in Cross-National Consumer Research." *Journal of Consumer Research* 25:78–90.

Sumner, William G. 1960. *Folkways: a Study of the Sociological Importance of Usages,*

Manners, Customs, Mores and Morals. New York: New American Library.

van de Schoot, Rens, Anouck Kluytmans, Lars Tummers, Peter Lugtig, Joop Hox, and Bengt 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.

van de Vijver, Fons J.R. 1998. "Towards a Theory of Bias and Equivalence." In *Zuma-Nachrichten Spezial Volume 3. Cross-Cultural Survey Equivalence*, edited by J. A. Harkness, 41–65. Mannheim, Germany: Zuma.

van der Veld, William, and Willem Saris. 2011. "Causes of Generalized Social Trust." In *Cross-Cultural Research: Methods and Applications*, edited by E. Davidov, P. Schmidt, and J. Billiet, 207-47. New York: Routledge

Vandenberg, Robert J. 2002. "Toward a Further Understanding of and Improvement in Measurement Invariance Methods and Procedures." *Organizational Research Methods* 5:139-58.

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* 3:4-70.

Vertovec, Steven. 2007. "Super-Diversity and its Implications." *Ethnic and Racial Studies* 30:1024-54.

Figure 1. A measurement model with a latent variable that measures attitudes toward immigration with three items (Item 1 – Item 3) and three measurement errors (e1-e3).

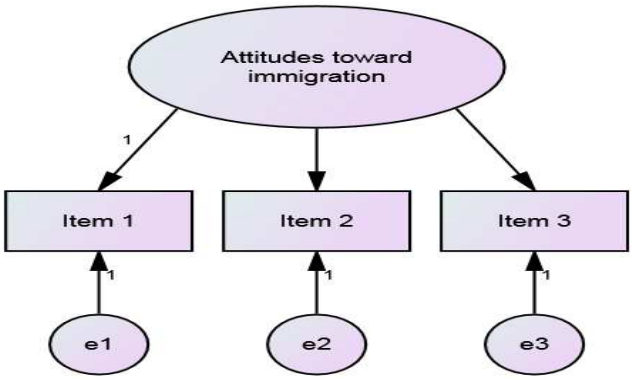

Table 1. Number of respondents (N) by country and ESS round

	Round 1 - 2002/03	Round 2 - 2004/05	Round 3 - 2006/07	Round 4 - 2008/09	Round 5 - 2010/11	Round 6 - 2012/13
1. Austria	2,053	2,074	2,236	1,987		
2. Belgium	1,739	1,619	1,645	1,586	1,516	1,606
4. Croatia				1,353	1,474	
6. Czech Republic	1,297	2,890		1,976	2,339	1,944
7. Denmark	1,422	1,415	1,403	1,510	1,475	1,536
8. Estonia		1,615	1,199	1,305	1,517	1,991
9. Finland	1,937	1,983	1,838	2,139	1,,813	2,103
10. France	1,353	1,670	1,791	1,911	1,573	
11. Germany	2,705	2,625	2,687	2,518	2,743	2,658
12. Greece	2,302	2,164		1,950	2,447	
13. Hungary	1,645	1,465	1,484	1,514	1,518	1,989
14. Iceland		554				707
15. Ireland	1,890	2,138	1,561	1,479	2,170	2,244
16. Israel	1,626			1,588	1,529	1,725
17. Italy	1,181	1,494				
18. Kosovo						1,222
19. Latvia			1,753	1,706		
20. Lithuania				1,916	1,592	
21. Luxembourg	1,069	1,147				
22. Netherlands	2,207	1,717	1,711	1,610	1,688	1,677
23. Norway	1,903	1,632	1,625	1,418	1,373	1,421
24. Poland	2,079	1,697	1,696	1,596	1,723	1,872
25. Portugal	1,421	1,932	2,078	2,229	2,004	2,019
26. Romania			2,130	2,088		
27. Russia			2,280	2,376	2,435	2,334
28. Slovakia		1,465	1,703	1,760	1,802	1,815
29. Slovenia	1,374	1,320	1,362	1,178	1,280	1,144
30. Spain	1,648	1,545	1,730	2,341	1,693	1,671
31. Sweden	1,785	1,762	1,710	1,616	1,324	1,613
32. Switzerland	1,696	1,748	1,464	1,392	1,155	1,157
33. Turkey		1,830		2,389		
34. Ukraine		1,763	1,759	1,654	1,717	
35. UK	1,860	1,724	2,158	2,106	2,151	2,020
Total	38,192	44,988	43,335	55,520	47,479	41,706

Notes. Empty cells denote that the country did not participate in the ESS in the respective round. The sample sizes represent individuals who were born in the country and are included in the analysis.

Table 2. Global fit measurements for the exact measurement equivalence test in each ESS round

	Chi2	df	RMSEA	SRMR	CFI
1st Round of ESS					
Configural	0.0	0	.000	.000	1.00
Metric	523.5	42	.083 [.076-.089]	.057	.993
Partial metric	200.5	21	.071 [.062-.080]	.029	.997
Partial scalar	465.7	42	.077 [.071-.084]	.037	.994
2nd Round of ESS					
Configural	0.0	0	.000	.000	1.00
Metric	890.3	50	.100 [.094-.106]	.075	.989
Partial metric	167.1	25	.058 [.050-.067]	.026	.998
Partial scalar	860.6	50	.098 [.092-.104]	.045	.989
3rd Round of ESS					
Configural	0.0	0	.000	.000	1.00
Metric	969.8	48	.107 [.101-.113]	.071	.987
Partial metric	282.1	24	.080 [.072-.082]	.032	.996
Partial scalar	1209.1	48	.120 [.114-.126]	.055	.984
4rd Round of ESS					
Configural	0.0	0	.000	.000	1.00
Metric	1501.2	60	.118 [.113-.123]	.083	.985
Partial metric	289.9	30	.071 [.063-.078]	.030	.997
Partial scalar	1283.0	60	.108 [.103-.114]	.050	.987
5th Round of ESS					
Configural	0.0	0	.000	.000	1.00
Metric	1108.9	52	.109 [.103-.115]	.074	.987
Partial metric	150.6	26	.053 [.045-061]	.022	.998
Partial scalar	1289.3	52	.118 [.112-.123]	.048	.985
6th Round of ESS					
Configural	0.0	0	.000	.000	1.00
Metric	964.6	46	.109 [.103-.115]	.076	.987
Partial metric	201.0	23	.068 [.059-.076]	.032	.998
Partial scalar	1353.1	46	.130 [.124-.136]	.059	.982

Notes. ESS = European Social Survey; Chi2 = chi-square; df = degrees of freedom; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; CFI = comparative fit index; Partial metric = released equality constraint on the factor loading of the item measuring whether respondents wish their country to allow many or few immigrants of the same race or ethnic group as the majority; partial scalar = released equality constraint on both the factor loading and intercept of that item in all countries.

Table 3. Countries where two or three intercepts were identified as misspecified by Jrule (with the criterion >.1).

ESS1	ESS2	ESS3	ESS4	ESS5	ESS6
9% countries	15% countries	40% countries	32% countries	37% countries	42% countries
Hungary	Estonia	Bulgaria	Bulgaria	Denmark	Cyprus
Israel	Portugal	Cyprus	Denmark	Estonia	Estonia
	Slovenia	Denmark	Estonia	Germany	Germany
	Ukraine	Estonia	Germany	Hungary	Hungary
		Hungary	Hungary	Israel	Iceland
		Latvia	Israel	Lithuania	Israel
		Russia	Latvia	Netherlands	Kosovo
		Spain	Lithuania	Spain	Netherlands
		Switzerland	Norway	Switzerland	Portugal
		Ukraine	Ukraine	Ukraine	Switzerland

Note. The table also reports the percentage of countries that did not reach partial scalar equivalence in the second row.

Table 4. Fit measurements for the approximate measurement equivalence model in each ESS round

	PPP	95% Confidence Interval
1st Round of ESS	.057	(-13.517) - (+108.288)
2nd Round of ESS	.422	(-53.57) - (+67.905)
3rd Round of ESS	.364	(-47.766) - (+68.527)
4rd Round of ESS	.220	(-44.291) - (+94.843)
5th Round of ESS	.340	(-52.088) - (+71.308)
6th Round of ESS	.320	(-45.631) - (+75.837)

Notes. 95% Credibility Interval = 95% Credibility Interval for the difference between the observed and the replicated chi-square values; PPP = the posterior predictive p-value

Table 5. Correlations of country rankings based on three methods (exact equivalence, approximate equivalence, and raw scores) in six ESS rounds (ESS1/ESS2/ESS3/ESS4/ESS5/ESS6)

	Exact (partial scalar model)	Approximate scalar model
Approximate scalar model	.995 / .998 / .993 / .988 / .992 / .973	
Raw scores	.954 / .971 / .970 / .956 / .971 / .963	.966 / .972 / .975 / .955 / .966 / .980

Note. ESS = European Social Survey