Survey Research Methods (2018)
Vol. 12, No. 1, pp. 15-27
doi:10.18148/srm/2018.v12i1.7212

© *European Survey Research Association*
ISSN 1864-3361
http://www.surveymethods.org

The cross-country measurement comparability in the immigration module of the European Social Survey 2014–15

Eldad Davidov
University of Cologne, Germany
and University of Zurich, Switzerland

Jan Cieciuch
University of Zurich, Switzerland
and Cardinal Wyszynski University
Warsaw, Poland

Peter Schmidt
University of Giessen, Germany
and Humboldt Research Fellow
Cardinal Wyszynski University
Warsaw, Poland

The 7th round of the European Social Survey (ESS) from 2014–15 includes a partial repetition of the immigration module from the first ESS wave (2002–03) with information on individual attitudes toward immigration and immigrants in both old and new immigration societies. The goal of the present study is to test whether and to what extent questions in the module are equivalent across ESS countries. We performed two types of measurement equivalence tests: exact and approximate. Whereas the exact approach requires that measurement parameters are exactly equal across groups, the approximate and newer approach suggests that it is sufficient that measurement parameters are approximately equal to allow a meaningful comparison across groups. Our findings suggest that two measurement scales, opposition toward immigration and realistic threat, are approximately invariant across most ESS countries and this allows the comparison of both associations with other theoretical constructs of interest and means.

Keywords: attitudes toward immigration; realistic threat; cross-country comparability; multigroup confirmatory factor analysis; exact and approximate measurement invariance; European Social Survey

1 Introduction

The European Social Survey (ESS) is a biennial survey that covers most West and East European countries and collects information about individuals' value orientations, attitudes, behavioral patterns and social structural position (see www.europeansocialsurvey.org). It includes a core part which is repeated every second year and a rotating part which covers diverse topics. The 7th round of the ESS (European Social Survey Round 7 Data, 2014) is a repeat module of the immigration module from the first ESS wave of 2002-03 (Preston, Bauer, Card, Dustmann, & Nazroo, 2001) that collected information on individual attitudes toward immigration and immigrants and diverse possible predictors of such attitudes in both old and new immigration societies. The immigration module in the ESS 2014-15 (Heath et al., 2014) partly replicates the immigration module included in the first wave of the ESS in 2002-03 as well as introduces new questions. A review of the module and its theoretical background can be found on the ESS website.[1]

The topic of immigration in the ESS has performed particularly well, and studies using ESS data on immigration have been highly cited (for an overview, see Kolarz et al., 2017). The new immigration module is very likely to be widely used by researchers as well. The present political and academic relevance of its topic area is pertinent today more than ever due to the ongoing large immigration flows into old and new immigration countries in Europe, the Great Recession of 2008 which resulted in an increase of anti-immigrant sentiments in many countries, and the continuing strength of radical right political parties focused on mobilizing public opposition to immigration. Therefore, the module is very likely to also serve as a research tool for a large number of comparative studies across countries.

Such cross-national studies have the potential to increase our knowledge about the prevalence of anti-immigrant sentiments and their antecedents across countries. However,

Contact information: Eldad Davidov, Institute of Sociology and Social Psychology, Albertus-Magnus-Platz, 50923 Cologne, Germany (e-mail: e.davidov@uni-koeln.de)

[1] http://www.europeansocialsurvey.org/methodology/questionnaire/ESS7_rotating_modules.html

such studies also face methodological challenges with respect to the comparability of the concepts used in different countries. Concepts in one country may not exist in another country, people may understand specific questions differently across countries, translations may be imprecise leading to biased scores, and people in various countries might use response scales differently when responding to survey questions (Cieciuch & Davidov, 2016; Cieciuch, Davidov, Oberski, & Algesheimer, 2015; Coromina & Davidov, 2013; Davidov, Meuleman, Cieciuch, Schmidt, & Billiet, 2014; Meuleman & Billiet, 2012). Therefore, before comparisons are conducted, it is crucial to test whether measurements of theoretical concepts are equivalent across countries.

In the current paper we are going to test for the measurement equivalence properties of three central concepts of the module included in the ESS 2014-15: attitudes toward immigration; (2) support of requiring qualifications from immigrants; and (3) realistic threat due to immigration. Findings of sufficient levels of equivalence across countries may provide important information for substantive researchers using these measures in the ESS: Such findings will allow meaningful conclusions on the similarities and differences in the occurrence and explanation of anti-immigrant sentiments across countries.

In the next section we will explain how measurement equivalence can be tested. We will use two methods to test for measurement equivalence: The exact (and most common but restrictive) approach and the rather new approximate (and more liberal) approach. Whereas the exact approach requires that measurement parameters are exactly equal across groups, the approximate, newer approach suggests that it is sufficient that measurement parameters are approximately equal to allow a meaningful comparison across groups. To illustrate, we will perform exact and approximate equivalence tests because recent studies suggest that the exact approach is too restrictive and that the approximate approach may be sufficient to guarantee comparability across countries (Cieciuch, Davidov, Schmidt, Algesheimer, & Schwartz, 2014; Davidov et al., 2015; Van de Schoot et al., 2013; Zercher, Schmidt, Cieciuch, & Davidov, 2015). After presenting these methods, we will describe the dataset we use and present the concepts included in the immigration module in the ESS on which we focus as well as the items that measure them. Next, we will present the data analysis and the findings. We will finalize with a summary and some concluding remarks.

2 Cross-country measurement comparability

Various techniques have been suggested in the literature (Brown, 2015; Davidov et al., 2014; Millsap, 2011) to test for so-called measurement equivalence (or measurement invariance). They demonstrate that differences in means or associations across groups such as countries or cultures are meaningful, only if it can be shown that specific measurement parameters are equal across groups. Establishing measurement equivalence can guarantee that the differences between groups cannot be traced back to methodological artefacts and that they reflect genuine cross-cultural differences.

Unfortunately, establishing measurement equivalence has often appeared to be a mammoth task (Byrne & van de Vijver, 2010). In most cases, high levels of equivalence cannot be established, especially when a large number of countries is involved. Recently, a new approach has been introduced that suggests testing for approximate rather than exact equality of measurement parameters (B. O. Muthén & Asparouhov, 2012, 2013; Van de Schoot et al., 2013). According to this approach, measurement parameters are allowed to vary slightly across countries. It is shown that these small differences do not threaten the meaningfulness of cross-country comparisons. Several recent empirical contributions have applied this approach in research and compared the results to those of the exact test (Cieciuch, Davidov, Algesheimer, & Schmidt, 2016; Cieciuch et al., 2014; Davidov et al., 2015; Zercher et al., 2015).

2.1 Exact measurement invariance

The exact measurement invariance approach distinguishes between three distinct and hierarchically ordered levels of measurement invariance: configural, metric, and scalar (Billiet, 2003; Millsap, 2011). The lowest level, configural invariance, simply guarantees that the same items do measure the same latent variables in each group. The second and higher level of exact measurement invariance is metric invariance. If it is given, factor loadings are equal across groups. This level of invariance suggests that the content and meaning of the latent variables are similar. If metric invariance is established, it would allow a meaningful comparison of associations between constructs (covariances and unstandardized regression coefficients) across groups (Ariely & Davidov, 2012; Davidov, Schmidt, & Schwartz, 2008; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). The third and highest level of exact measurement invariance is scalar equivalence (Horn & McArdle, 1992; Steenkamp & Baumgartner, 1998; Vandenberg, 2002; Vandenberg & Lance, 2000). It requires that not only the factor loadings but also the item intercepts are equal across groups (Meredith, 1993). This level of equivalence, if established, would also allow meaningful comparisons of latent variable means across groups. Multigroup confirmatory factor analysis (MGCFA) is typically used for testing for exact measurement equivalence (Bollen, 1989; Brown, 2015; Jöreskog, 1971).

There are various ways to estimate whether exact measurement equivalence is supported by the data. Whereas the chi-square difference test has been considered in the last two decades to be too strict (often leading to the rejection of measurement equivalence, even when differences between measurement parameters are rather small or trivial; see, e.g., Chen, 2007; Cheung & Rensvold, 2002), alternative and more liberal criteria have become more common. The most frequently used ones are probably the approaches proposed by Cheung and Rensvold (2002) and Chen (2007). They suggest determining whether global fit measures such as the comparative fit index (CFI) and/or the root mean square error of approximation (RMSEA) considerably deteriorate when moving from a less to a more restrictive model (i.e., from configural to metric or from metric to scalar equivalence). A difference of 0.01 or less in CFI supplemented by a difference of 0.015 or less in RMSEA suggest that the more restrictive model may be accepted for samples larger than 300 (for details see, e.g., Chen, 2007; for an application see, e.g., Byrne and Stewart, 2006).

Particularly scalar equivalence is difficult to establish. This may be due to the high power of the test (Saris, Satorra, & van der Veld, 2009), but also due to social desirability bias, acquiescence, or other response pattern differences across cultural groups (e.g., Billiet, 2003; Oberski, 2014). In such cases one may resort to relying on partial rather than full invariance. Byrne, Shavelson, and Muthén (1989) suggest that two items with equal factor loading and intercepts are sufficient to establish partial (rather than full) invariance (see also Steenkamp & Baumgartner, 1998). However, it could also very well be the case that the exact test is too strict, and maybe the approximate rather than exact equality of item intercepts and factor loadings would be sufficient for conducting meaningful comparisons.

2.2 Approximate measurement invariance

Recently, B. O. Muthén and Asparouhov (2013) and Van de Schoot et al. (2013) proposed an alternative approach to test for measurement invariance by applying approximate Bayesian measurement invariance testing. The basic idea of the method is to replace exact equality constraints in factor loadings and intercepts with approximate equality constraints. The Bayesian approach allows incorporating approximate equality constraints by using Bayesian priors with a mean of zero and a small variance. Thus, following this approach we would allow the introduction of some uncertainty by specifying a small variance (such as 0.01, 0.05, or 0.1) around the difference in factor loadings or intercepts (Van de Schoot et al., 2013).

In the traditional exact approach we constrain differences between factor loadings or intercepts between groups to be zero. However, in the approximate approach we constrain differences between these parameters to be approximately zero (Kruschke, Aguinis, & Joo, 2012; Levy & Choi, 2013; B. O. Muthén & Asparouhov, 2013). This is operationalized by constraining the mean difference between parameters to be zero, and its variability to be larger than zero but small. Simulation studies by Van de Schoot et al. (2013) suggest that "small" variations of 0.05 may be allowed without risking invalid conclusions.

Two fit measures are typically used to determine whether approximate equivalence is given or not (Gelman, 2003, 2013; Levy, 2011). Similar to the exact approach, they provide information on whether and to what extent measurement parameter deviations across groups in the given data are larger than those allowed by the researcher in the prior variance distribution.

1. The first fit measure is the posterior predictive probability value (ppp). The ppp is computed by comparing the discrepancy between the model and the observed data and the discrepancy between the model and the posterior predicted data (Levy & Choi, 2013; Zercher et al., 2015). B. O. Muthén and Asparouhov (2012) and Van de Schoot et al. (2013) suggest that in order to determine whether approximate equivalence is present in the data, the ppp value should be non-significant. Furthermore, a ppp value of 0.50 and higher suggests that the model fits the data very well.

2. The second fit measure is the credibility interval (CI). The credibility interval informs about the difference between the observed and the replicated chi-square values. B. O. Muthén and Asparouhov (2012) and Van de Schoot et al. (2013) suggest that in order to determine whether approximate equivalence is present in the data, the credibility interval should contain zero. Before applying the two methods to analyze the measurement invariance properties of measurements in the ESS immigration module, in the next section we describe the data and measures we analyze.

3 THE CURRENT STUDY

3.1 Data

The data we analyze in the study were retrieved from www.europeansocialsurvey.org. We include 15 countries in the analysis: Austria, Belgium, Switzerland, Czech Republic, Germany, Denmark, Estonia, Finland, France, Ireland, Netherlands, Norway, Poland, Sweden, and Slovenia. Table 1 presents the sample size in each country, the mean age, and the percentage of females among the respondents in each country. The ESS website includes further information and documentation about sampling procedures and the questionnaires.

3.2 Measurements

The proposal for the repeat module (Heath et al., 2014) differentiates between concepts and constructs on the one hand and items on the other hand, defines each of the constructs in the study, and makes specific suggestions which items should measure specific constructs. These suggestions are based on previous studies from the immigration literature. We follow these suggestions in the current study. The goal of the current study is to empirically examine whether

Table 1
Number of respondents in each country with mean age and percentage of females in each sample

Country	N	Age		Female
		Mean	Std. Dev.	%
Austria	1, 795	49.22	18.06	52.5
Belgium	1, 769	46.94	18.97	49.3
Czech Republic	2, 148	46.80	17.06	52.5
Denmark	1, 502	48.13	18.94	48.1
Estonia	2, 051	50.32	19.08	59.3
Finland	2, 087	51.31	19.07	50.8
France	1, 917	49.98	18.74	52.4
Germany	3, 045	49.90	18.39	49.3
Ireland	2, 390	49.39	18.19	53.9
Netherlands	1, 919	50.74	18.25	55.2
Norway	1, 436	46.77	18.68	46.8
Poland	1, 615	47.30	18.80	54.2
Slovenia	1, 224	49.58	18.65	54.0
Sweden	1, 791	49.70	19.90	50.1
Switzerland	1, 532	47.36	18.23	50.0
Total sample	28, 221			

Data source: ESS 2014-15 (7th round)

measures in the 2014/15 immigration module are equivalent across ESS countries. Since such a test requires multiple indicators to measure the concepts of interest, we focus on three latent variables: opposition toward immigration, qualification for entry or exclusion, and realistic threat. We did not include other scales in the module in the measurement invariance test because they were measured by only two items or by a single indicator. When only two items were available, it was not possible to control for every type of nonrandom measurement error (Bollen, 1989). When only one indicator was available, no control for random and nonrandom measurement errors was possible. The construct opposition toward immigration (Allowance) was measured by four questions asking to what extent respondents think [country] should allow people from other countries to come and live in [country]. The question referred to four more specific groups: people of a different race, Jews, Muslims, and Gypsies. Response categories ranged from 1 (allow many to come and live here) to 4 (allow none).[2] The construct Qualification for entry or exclusion (Conditions) was measured by six questions inquiring how important respondents think each of these things should be in deciding whether someone born, brought up, and living outside [country] should be able to come and live in [country]: having good educational qualifications; being able to speak [country's official language(s)]; coming from a Christian background; being white; having work skills that [country] needs; and being committed to the way of life in [country]. Response categories ranged from 1 (extremely unimportant) to 10 (extremely important). The construct realistic (economic and security) threat due to immigration (RT) was measured by four questions inquiring whether respondents agree that immigrants take jobs away, take out more than they put in, make crime problems worse, and are bad for the country's economy. Response categories ranged from 0 (highest threat) to 10 (lowest threat). Table 2 presents the items and the constructs they measure, the item formulation, and the response categories with their labels.

We proceed with the analysis of their measurement invariance properties in two steps. In the first step we test the proposed measurement models in each country separately. We rely on the CFI and RMSEA global fit measures to assess whether the model was supported by the data in each country. CFI values higher than 0.95 combined with RMSEA values lower than 0.08 indicate a good fit (West, Taylor, & Wu, 2012). We also report the χ^2 and the number of degree of freedom (df) for each model but do not rely on them to determine the fit, because the χ^2 is considered to be too strict leading to a rejection of the model too often. In the second step we use the baseline model that was supported by the data in the first step and conduct the measurement invariance tests. The percentage of missing values ranged between 4.6% (for the two RT items measuring whether immigrants increase crime rates and whether they put in more than they take out) and 0.7% (for the item measuring whether immigrants should be able to speak the country's official language). To address this problem we applied the full information maximum likelihood (FIML) procedure, which is implemented in the software package Mplus that we use for the analyses (L. Muthén & Muthén, 2015). We run the tests using two approaches, the exact one and the approximate one, and report the results for each procedure.

4 RESULTS

4.1 Opposition toward immigration (Allowance)

The model of the latent variable Allowance, with its four indicators, did not fit the data well in all countries. We improved the model fit by adding an error correlation between

[2] Four more general questions measure opposition toward immigration and the willingness to allow different immigrant groups into the country (e.g., of a different race or ethnic group, or from poorer countries in or outside of Europe). However, these questions loaded on a separate factor and were likely to represent a separate dimension of opposition toward immigration. Three of these questions were part of the core part of the ESS questionnaire and were therefore not included in the present study. Davidov et al. (2015) tested for exact and approximate measurement invariance of these items and found cross-country approximate measurement invariance for the items. Finally, a question in the module inquiring whether immigrants make the country a worse or a better place to live did not load on the same factor and is likely to represent a separate dimension of opposition toward immigration.

Table 2
Constructs and items measuring them, item formulations, and response scales

Allowance Allowing for immigrants belonging to ethnic groups different than the majority population to come into the country.

Question: To what extent to you think [country] should allow . . . people from other countries to come and live in [country]?

Variant	Abbreviation	ESS item name
different race	Race	imdfetn
Jewish	Jewish	aljewlv
Muslims	Muslim	almuslv
Gypsies	Gypsies	algyplv

Response scale: (1) allow many to come and live here; (2) allow some; (3) allow a few; (4) allow none

Conditions Qualification for entry.

Question: How important do you think each of these things should be in deciding whether someone born, brought up, and living outside [country] should be able to come and live here. Firstly, how important should it be for them to . . .

Variant	Abbreviation	ESS item name
. . . have good educational qualifications?	Eduqual	qfimedu
. . . be able to speak [country's official language(s)]?	Language	qfimlng
. . . come from Christian background?	Christian	qfimchr
. . . be white?	White	qfimwht
. . . have work skills that [country] needs?	Workskills	qfimwsk
. . . be committed to the way of life in [country]?	Wayoflife	qfimcmt

Response scale: (1) extremely unimportant; (10) extremely important

RT Realistic threat

Question	Abbreviation	ESS item name	Resp. Scale
Would you say that people who come to live here generally take jobs away from workers in [country], or generally help to create new jobs?	Jobs	imtcjob	(0) take jobs away; (10) create new jobs
Would you say it is generally bad or good for [country]'s economy that people come to live here from other countries?	Economy	imbgeco	(0) bad for the economy (10) good for the economy
Are [country]'s crime problems made worse or better by people coming to live here from other countries?	Crime	imwbcrm	(0) crime problems made worse; (10) crime problems made better
Most people who come to live here work and pay taxes. They also use health and welfare services. On balance, do you think people who come here take out more than they put in or put in more than they take out?	Putmore	imbleco	(0) generally take out more (10) generally put in more

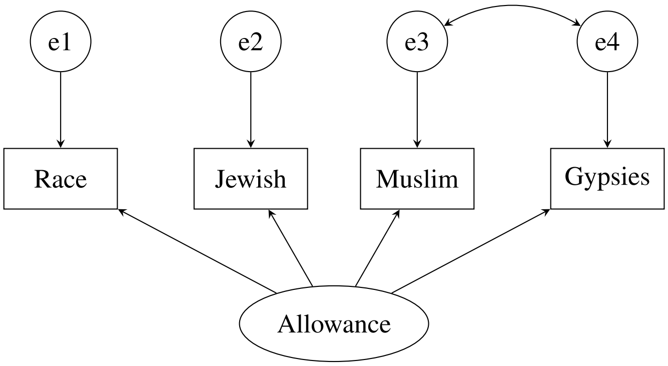

Figure 1. The latent variable Allowance. Item abbreviations are presented in Table 2

Table 3
Global fit measures for the measurement model of Allowance in each country (single-country CFAs)

		RMSEA			
Country	$\chi^{2\text{a}}$	Est.	Lower	Upper	CFI
Austria	5.82	0.052	0.018	0.096	1.00
Belgium	4.10	0.042	0.006	0.087	1.00
Czech Republic	3.59	0.035	0.000	0.076	1.00
Denmark	4.21	0.046	0.008	0.096	1.00
Estonia	1.36	0.013	0.000	0.062	1.00
Finland	2.74	0.029	0.000	0.072	1.00
France	89.48	0.216	0.179	0.255	1.00
Germany	4.41	0.034	0.007	0.068	1.00
Ireland	21.11	0.092	0.060	0.128	1.00
Netherlands	0.04	0.000	0.000	0.051	1.00
Norway	6.64	0.063	0.025	0.111	1.00
Poland	0.31	0.000	0.000	0.054	1.00
Slovenia	0.38	0.083	0.041	0.136	1.00
Sweden	0.22	0.000	0.000	0.049	1.00
Switzerland	0.06	0.000	0.000	0.043	1.00

[a] df = 1

two items: allowing for Muslims and allowing for Gypsies. It seems that the two items that measure attitudes toward the most rejected groups in Europe (Heath & Ford, 2016) are more similar to each other than the other items[3]. This model is presented in Figure 1.

The fit indices of the model in each country are presented in Table 3. Based on the CFI fit measure, the model fit the data well in all 15 countries. However, based on the RMSEA fit measure the model fit the data well in all countries but three: France, Ireland, and Slovenia. Therefore, in the following models we excluded these countries from the analysis.

In the next step we tested the measurement invariance properties of the model across all countries excluding the three countries where the measurement model did not fit the data well.[4] The analyses were conducted using the exact and approximate approaches. Results are presented in Table 4 (exact approach) and Table 5 (approximate approach).

In the exact approach, based on the CFI fit measure, one could conclude that both metric and scalar measurement invariance were supported across the 12 countries. However, such a conclusion is not possible based on the RMSEA fit measure. This index instead suggests a lack of metric and scalar measurement invariance for both the full and partial invariance models. Such an inconsistent pattern, in which one fit measure indicates measurement invariance whereas the other does not, may suggest that the violation of measurement invariance is not severe.[5] This expectation is supported by the results of the test for approximate measurement invariance. Given that the ppp was not significant and the 95% CI contained zero, approximate measurement invariance across 12 countries was supported by the data.[6]

[3] Researchers examining the concept Allowance across countries using these items are thus advised to allow for this error correlation to avoid biased estimates (see Brown, 2015).

[4] It could well be the case that an additional error correlation will further improve the fit of the model in France, Ireland, and Slovenia. We preferred to avoid it to be able to operate with a simpler model. Researchers interested in any of these three countries may consider using a slightly modified model and test its invariance properties with their other countries of interest.

[5] Chen (2007) suggests that CFI and RMSEA may be similarly sensitive to violations of invariance under different conditions. Thus, when invariance is not given, one would expect both of them to perform badly. As only the RMSEA did not perform very well in this case, it is likely that the degree of noninvariance is not severe enough to affect both fit measures.

[6] Statistical support for measurement invariance is a necessary but not a sufficient condition for a similar understanding of the concepts. Cognitive interviews offer a supplementary tool to assess the equivalence of meaning of the instruments across countries (Meitinger, 2017).

Table 4
Measurement invariance test of Allowance – the exact approach

			RMSEA			
	χ^2	df	Est.	Lower	Upper	CFI
Multigroup confirmatory factor analysis across 12 countries[a]						
Configural	53	12	0.042	0.031	0.054	1.00
Metric	955	45	0.103	0.098	0.109	0.97
Scalar	4186	166	0.113	0.110	0.116	0.98
Partial Metric and Scalar[b]	2133	78	0.118	0.114	0.122	0.99

[a] Without France, Ireland, Slovenia [b] Released loadings and thresholds in the items race (allowing immigrants of a different race) and Jewish (allowing Jews). For these items, the violation of the equality constraints for the measurement parameters was the strongest.

Table 5
Measurement invariance test of Allowance – the approximate approach[a]

		95% CI	
Analysis	PPP	Lower	Upper
Across 12 countries[b]	0.178	−29.967	73.803

[a] Prior variance = 0.05
[b] Without France, Ireland, Slovenia

Table 6
Measurement model (single CFA) of Conditions in each country

		RMSEA			
Country	χ^{2a}	Est.	Lower	Upper	CFI
Austria	140	0.103	0.088	0.118	0.96
Belgium	65	0.068	0.054	0.084	0.98
Czech Republic	184	0.109	0.095	0.123	0.95
Denmark	39	0.055	0.038	0.072	0.99
Estonia	126	0.091	0.078	0.105	0.96
Finland	133	0.093	0.080	0.107	0.97
France	222	0.127	0.113	0.141	0.94
Germany	139	0.079	0.068	0.091	0.98
Ireland	199	0.107	0.095	0.120	0.95
Netherlands	51	0.057	0.043	0.073	0.93
Norway	34	0.052	0.035	0.070	0.99
Poland	142	0.110	0.095	0.126	0.95
Slovenia	87	0.097	0.079	0.116	0.95
Sweden	64	0.067	0.053	0.083	0.99
Switzerland	82	0.083	0.068	0.100	0.97

[a] df = 7

4.2 Qualification for entry or exclusion

The latent variable consisting of six indicators measuring support of qualification requirements for entry of immigrants into the country is presented in Figure 2. In the single country CFAs, this model did not fit the data well in all 15 countries. We could improve the model by introducing two error correlations: between the items “being white” and “coming from a Christian background” and between “educational qualifications” and “work skills that country needs”. Respondents’ support of education and work skills as entry qualifications are more strongly associated with each other than with the support of other qualifications. Both education and work skills are attributes which are relevant for the successful integration into the labor market of the host society. Similarly, respondents’ support of requiring immigrants to be white and from a Christian background are also more strongly associated with each other than with respondents’ support of requiring other qualifications. Both are attributes immigrants are born with (being Christian may be also acquired).[7]

The global fit indices of the model in each country are presented in Table 6. Based on both the CFI and RMSEA global fit measures, the model is acceptable only in seven countries: Belgium, Germany, Denmark, Netherlands, Norway, Sweden, and Switzerland (CFI > 0.90, RMSEA < 0.08, Switzerland was an equivocal case). In other words, even configural invariance was not supported by the data for the other countries because the items did not measure the concepts across all countries in a consistent way.

In the next step we tested the measurement invariance of the model presented in Figure 2 across the seven countries where the measurement model was acceptable. Both analyses were performed using the exact and approximate approaches. Results are presented in Table 7 (exact approach) and 8 (approximate approach).

Based on the global fit measures of the analyses presented in Tables 7 and 8, one can conclude that neither (full or partial) exact nor approximate scalar invariance was supported

[7] Researchers using this scale are advised to introduce these error correlations to avoid biased estimates of the construct’s variance and covariance with other theoretical constructs of interest (Brown, 2015).

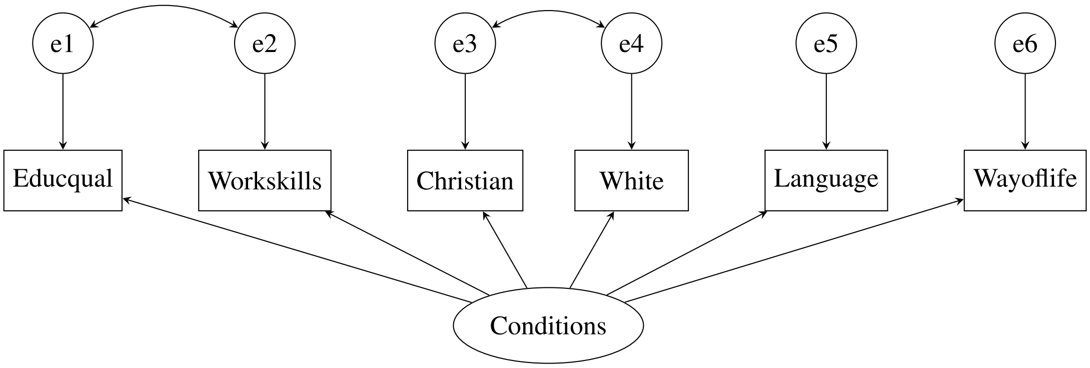

Figure 2. The latent variable Conditions. Item abbreviations are presented in Table 2

Table 7
Measurement invariance test of Conditions – the exact approach

			RMSEA			
	χ^2	df	Est.	Lower	Upper	CFI
Multigroup confirmatory factor analysis across 7 countries[a]						
Configural	473	49	0.068	0.063	0.074	0.98
Metric	924	79	0.076	0.072	0.080	0.96
Scalar	3321	109	0.126	0.122	0.130	0.85
Partial Metric and Scalar[b]	1422	72	0.101	0.096	0.105	0.94

[a] Belgium, Germany, Denmark, Netherlands, Norway, Sweden, Switzerland

[b] Released loadings and intercepts for the items language (immigrants should be able to speak the country's language), Christian (immigrants should come from a Christian background), and wayoflife (immigrants should be committed to the way of life in the country). For these items, the violation of the equality constraints for the measurement parameters was the strongest.

Table 8
Measurement invariance test of Conditions – the approximate approach[a]

		95% CI	
Analysis	PPP	Lower	Upper
Across 7 countries[b]	0.000	412	526

[a] Prior variance = 0.05

[b] Belgium, Germany, Denmark, Netherlands, Norway, Sweden, Switzerland

by the data even across this subset of countries where the measurement model presented in Figure 2 fit the data relatively well. Metric invariance was, however, supported across this set of countries. Indeed, the topic of defining specific qualifications as conditions for entry of immigrants into the country is highly debated, and it could well be the case that the items measuring support of these qualifications are susceptible to different levels of response bias across various European countries. Nonetheless, Conditions may be comparable across a different and specific subset of countries that we did not examine.

4.3 Realistic threat

The latent variable with four indicators measuring realistic threat due to immigrants is presented in Figure 3. To achieve a better fit for the model it was necessary to include an error correlation between the two items relevant for the economy.[8] The two items measured agreement with the statements that immigrants take away jobs and are bad for the economy. This model fit the data well in all countries except Finland (where the CFI displayed a good fit but the RMSEA did not). Table 9 presents the global fit coefficients in each single-country analysis.

In the next step we tested the measurement invariance properties of the model presented in Figure 3 across all countries except Finland and across all countries except Finland and Sweden[9]. Both analyses were conducted using the exact

[8] Researchers using this scale are advised to introduce this error correlation to avoid biased estimates of the construct's variance and covariance with other theoretical constructs of interest (Brown, 2015).

[9] The model fits the data very well in Sweden as evidenced in

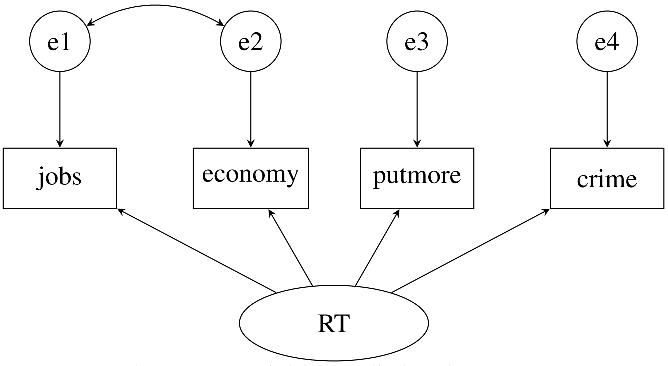

Figure 3. The latent variable Realistic Threat (RT). Item abbreviations are presented in Table 2

Table 9
Measurement model (single CFA) of RT in each country

Country	χ^{2a}	RMSEA			CFI
		Est.	Lower	Upper	
Austria	0.03	0.000	0.000	0.032	1.00
Belgium	6.09	0.054	0.019	0.098	1.00
Czech Republic	5.07	0.044	0.013	0.084	1.00
Denmark	0.82	0.000	0.000	0.066	1.00
Estonia	6.95	0.054	0.022	0.095	1.00
Finland	20.98	0.098	0.064	0.137	1.00
France	11.39	0.074	0.040	0.115	0.99
Germany	5.62	0.039	0.013	0.073	1.00
Ireland	0.02	0.000	0.000	0.024	1.00
Netherlands	0.01	0.000	0.000	0.022	1.00
Norway	2.36	0.031	0.000	0.084	1.00
Poland	0.72	0.000	0.000	0.062	1.00
Slovenia	1.18	0.012	0.000	0.079	1.00
Sweden	12.35	0.080	0.044	0.122	0.99
Switzerland	1.32	0.015	0.000	0.071	1.00

[a] df = 1

and approximate approaches. Results are presented in Table 10 (exact approach) and Table 11 (approximate approach).

Based on the fit measures presented in Tables 10 and 11, we can conclude that neither full nor partial exact scalar measurement invariance was established across the 14 countries. Also, approximate invariance was not given in the data because the CI did not contain a zero. However, after dropping Sweden from the analysis, approximate (but not partial exact) scalar invariance could be supported by the data.

5 SUMMARY AND DISCUSSION

The increasing availability of large-scale cross-cultural and cross-country surveys in the last decades has significantly increased the possibilities for researchers to conduct comparative studies. However, they have also considerably increased the risk encountered by researchers of drawing incorrect conclusions if the measurements in such studies are not equivalent across groups. As a countermeasure, the methodological literature on cross-cultural analysis has recommended testing for measurement equivalence to guarantee that differences across groups are due to substantive true differences and not a result of methodological artefacts. This recommendation has been increasingly applied in the last decade by various researchers who have examined the measurement equivalence properties of various scales (for an overview, see Davidov et al., 2014). Unfortunately, a new problem has come up: Many scales have failed to display high levels of equivalence.

Approximate equivalence has been proposed in the literature to circumvent this problem. According to this approach, measurement parameters are allowed to vary a little across groups, and several studies have shown that these small differences do not threaten the meaningfulness of cross-group comparisons (e.g., Davidov et al., 2015). These studies have furthermore demonstrated that approximate equivalence may also be given when exact equivalence is rejected by the data. However, as Davidov et al. (2015) clearly point out, the procedure “does not do magic”. When measurement parameters are “too different”, even this more liberal test will fail to establish approximate equivalence. These are good news: After all, we want the test to provide us with information about (approximate) equivalence only when deviations are small enough not to distort the interpretation of substantive differences across countries. It helps us to be more flexible in the measurement equivalence test while still allowing a meaningful comparative analysis. The introduction of this approach constitutes a further step in fulfilling one of Roger Jowell’s golden rules to use sound survey methodology for a meaningful interpretation in comparative research (Jowell,

Table 10. However, the violations of metric and scalar invariance are stronger in Sweden than in the other countries. Dropping Sweden, as shown below, assisted us in achieving approximate (but not exact) invariance.

Table 10
Measurement invariance test of RT – the exact approach

			RMSEA			
	χ^2	df	Est.	Lower	Upper	CFI
Multigroup confirmatory factor analysis across 14 countries[a]						
Configural	54	14	0.039	0.028	0.050	1.00
Metric	539	53	0.070	0.065	0.076	0.98
Scalar	4433	92	0.159	0.155	0.163	0.82
Multigroup confirmatory factor analysis across 13 countries[b]						
Configural	42	13	0.034	0.023	0.046	1.00
Metric	473	49	0.068	0.063	0.074	0.98
Scalar	4041	85	0.158	0.154	0.162	0.82
Partial Metric and Scalar[c]	1627	45	0.137	0.132	0.143	0.93

[a] Without Finland [b] Without Finland and Sweden

[c] Released loadings and intercept for the items crime (immigrants make crime problems worse) and economy (immigrants are generally good for the economy). For these items, the violation of the equality constraints for the measurement parameters was the strongest. Sweden was dropped from this analysis because its violations of invariance were the strongest of all countries.

Table 11
Measurement invariance test of RT – the approximate approach[a]

		95% CI	
Analysis	PPP	Lower	Upper
Across 14 countries[b]	0.017	4.43	118.19
Across 13 countries[c]	0.041	−6.30	104.30

[a] Prior variance = 0.05 [b] without Finland

[c] without Finland and Sweden

Roberts, Fitzgerald, & Gillian, 2007).[10]

The 7th round of the ESS included a repeat module with questions measuring attitudes toward immigration and immigrants. The goal of the present study was to test whether these measures are equivalent across ESS countries. We performed tests of exact and approximate measurement invariance on the following scales which were measured by multiple indicators: opposition toward immigration (Allowance), qualifications for entry or exclusion (Conditions), and realistic threat due to immigration (RT).

Results provided empirical support of scalar approximate (but not exact) invariance for Allowance and RT (across most countries). Thus, results suggest that scores of the two constructs and their association with other theoretical constructs of interest may be compared across most ESS countries with confidence. However, results did not support exact or approximate scalar invariance for Conditions across the countries we studied but only metric invariance across a limited set of countries. Whereas the measurements of opposition toward immigration or of realistic threat due to immigrants seem to operate rather similarly across European countries, this is not the case for the more concrete measurements of the construct Conditions. These questions measure the support of requiring specific qualifications from immigrants. Because the social and economic needs or expectations of different countries from immigrants may vary considerably, it could very well be the case that also the measurements of support for qualifications that fulfill these needs and their respective response patterns vary accordingly. Consequently, comparing scores of this construct across ESS countries should be done with much more caution or avoided when scores are not invariant. However, since metric invariance was supported by the data across a subset of countries, comparing associations (unstandardized regression coefficients or covariances) between Conditions and other theoretical constructs of interest may be meaningful across these countries.[11]

[10] For an alternative and new approach to address measurement equivalence, the alignment optimization method, see also Asparouhov and Muthén (2014); for applications, see, for example, Cieciuch, Davidov, and Schmidt (2018) and Munck, Barber, and Torney-Purta (Online first). Nonetheless, findings of noninvariance may be a useful source of information on country differences (Davidov et al., 2016; Jak, Oort, & Dolan, 2014).

[11] Meuleman and Billiet (2012) tested for the measurement invariance properties of four constructs in the ESS 2002/3 immigration module. Whereas Allowance (which was measured in their study by different items than in our study) reached measurement invariance across all countries, partial scalar invariance for the Conditions scale could be reached only for a subset of countries. The authors also tested the measurement invariance properties of the eco-

The immigration module in the ESS includes additional scales of which we did not examine the measurement invariance properties. The reason is that these scales were not measured by a sufficient number of indicators to allow us to perform an invariance test (i.e. they were measured by either two or only one single indicator). Therefore, using these scales in comparative studies should be done with caution.

Below we would like to provide some general recommendations and concluding remarks for applied researchers interested in utilizing the scales presented in this study in a comparative perspective. First, while the analyses performed in the current study suggest which countries display measurement invariance, they cannot provide information as to whether groups within these countries are measurement invariant. Researchers who are interested in comparing social groups (e.g., males and females, old and younger respondents, immigrants and natives, respondents from different geographical regions) within these countries should perform similar analyses across these groups to assess exact or approximate measurement invariance for the scales (for illustrations across regional groups, see, e.g., Sarrasin, Green, Berchtold, and Davidov, 2012; Siegers and Davidov, 2010; for an illustration across education groups, see, e.g., Steinmetz, Schmidt, Tina-Booh, Wieczorek, and Schwartz, 2010. Findings of exact or approximate invariance of the measures will allow carrying out further substantive analyses with their scores and drawing meaningful conclusions on similarities and differences across groups with more confidence.[12] Second, even when measurement invariance is not given, it may still be possible to compare specific countries meaningfully. For example, although the construct Conditions appeared to be noninvariant across the countries we analyzed, it may be scalar invariant across a different and specific subset of countries. Researchers interested in the mean comparison of Conditions across specific countries should perform similar analyses across these countries to assess exact or approximate measurement invariance for the scales (for alternative procedures to assess measurement invariance, see Davidov et al., 2014; Davidov, Schmidt, J., & Meuleman, 2018). Finally, the findings suggest that Allowance and RT may be used for cross-country comparisons meaningfully, as they displayed approximate scalar invariance across most ESS countries in the study.[13]

Acknowledgments

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

References

Ariely, G. & Davidov, E. (2012). Can we rate public support for democracy in a comparable way? Cross-national equivalence of democratic attitudes in the World Value Survey. *Social Indicators Research*, *104*(2), 271–286.

Asparouhov, T. & Muthén, B. O. (2014). Multi-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*, 1–14.

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

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

Braun, M. & Johnson, T. P. (2018). How should immigrants adapt to their country of residence? A mixed methods approach to evaluate the international applicability of a question from the German General Social Survey (ALLBUS). In E. Davidov, P. Schmidt, J. Billiet, & B. Meuleman (Eds.), *Cross-cultural analysis: methods and applications* (2nd ed., pp. 615–632). New York: Routledge.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York: Guilford Press.

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

Byrne, B. M. & Stewart, S. M. (2006). The macs approach to testing for multigroup invariance of a second-order structure: a walk through the process. *Structural Equation Modeling*, *13*(2), 287–321.

nomic threat scale which shares two of the items of the RT scale in our study. Partial scalar invariance for this scale could be reached only for a subset of countries in their study. Thus, our findings are essentially in line with the Meuleman and Billiet (2012) study. However, it should be noted that the 1st ESS immigration module from 2002/3 did not include precisely the same set of items on attitudes toward immigration as those in the repeat module. Indeed, the repeat module in Round 7 is only a partial repetition and includes several new items. In addition, the set of countries across the 1st and 7th ESS modules is not identical.

[12] MIMIC models (multiple indicators multiple causes models; see Schumacker & Lomax, 2010) could offer an additional tool to assess whether items operate differently across specific groups. If they operate similarly, scores of the latent variables may be compared across these groups meaningfully.

[13] Yet it should be noted that tests of measurement invariance provide necessary but not sufficient conditions of comparability. Supplementary probing techniques may provide further evidence on whether and to what extent items are comparable across countries (for a review, see Braun & Johnson, 2018; Meitinger, 2017).

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

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, *14*(3), 464–504.

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

Cieciuch, J. & Davidov, E. (2016). Establishing measurement invariance across online and offline samples. a tutorial with the software packages Amos and Mplus. *Studia Psychologica*, *15*(2), 83–99.

Cieciuch, J., Davidov, E., Algesheimer, R., & Schmidt, P. (2016). Testing for approximate measurement invariance of human values in the European Social Survey. *Sociological Methods & Research*.

Cieciuch, J., Davidov, E., Oberski, D. L., & Algesheimer, R. (2015). Testing for measurement invariance by detecting local misspecification and an illustration across online and paper-and-pencil samples. *European Political Science*, *14*, 521–538.

Cieciuch, J., Davidov, E., & Schmidt, P. (2018). Alignment optimization: estimation of the most trustworthy means in cross-cultural studies even in the presence of noninvariance. In E. Davidov, P. Schmidt, J. Billiet, & B. Meuleman (Eds.), *Cross-cultural analysis: methods and applications (2nd edition)* (pp. 571–593). New York: Routledge.

Cieciuch, J., Davidov, E., Schmidt, P., Algesheimer, R., & Schwartz, S. H. (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), 1–10.

Coromina, L. & Davidov, E. (2013). Evaluating measurement invariance for social and political trust in Western Europe over four measurement time points (2002–2008). *ASK Research & Methods*, *22*(1), 37–54.

Davidov, E., Cieciuch, J., Meuleman, B., Schmidt, P., Algesheimer, R., & Hausherr, M. (2015). The comparability of measurements of attitudes toward immigration in the European Social Survey: exact versus approximate measurement equivalence. *Public Opinion Quarterly*, *79*, 244–266.

Davidov, E., Dülmer, H., Cieciuch, J., Kuntz, A., Seddig, D., & Schmidt, P. (2016). Explaining measurement nonequivalence using multilevel structural equation modeling. *Sociological Methods & Research*.

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology*, *40*, 55–75.

Davidov, E., Schmidt, P., J., B., & Meuleman, B. ((2018). *Cross-cultural analysis: methods and applications*. New York: Routledge.

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

European Social Survey Round 7 Data. (2014). Data file edition 2.1. NSD – Norwegian Centre for Research Data, Norway – Data Archive and distributor of ESS data for ESS ERIC. Retrieved from http://www.europeansocialsurvey.org/data/download.html?r=7

Gelman, A. (2003). A bayesian formulation of explanatory data analysis and goodness of fit testing. *International Statistics Review*, *71*, 369–382.

Gelman, A. (2013). Two simple examples for understanding posterior p-values whose distributions are far from uniform. *Electronic Journal of Statistics*, *7*, 2595.

Heath, A. & Ford, R. (2016). *How do europeans differ in their attitudes to immigration?* Paper presented at the 3rd ESS conference, July 2016, in Lausanne, Switzerland.

Heath, A., Schmidt, P., Green, E., Ramos, A., Davidov, E., & Ford, R. (2014). European Social Survey round 7 repeat module proposal. Attitudes towards immigration and their antecedents. Retrieved from http://www.europeansocialsurvey.org/methodology/questionnaire/ESS7%5C_rotating%5C_module%5C_immigration.html

Horn, J. L. & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*(3), 117–144.

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

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*(4), 409–426.

Jowell, R., Roberts, C., Fitzgerald, R., & Gillian, E. (Eds.). (2007). *Measuring attitudes cross-nationally: lessons from the European Social Survey.* London: Sage.

Kolarz, P., Angelis, J., Krčál, A., Simmonds, P., Traag, V., & Wain, M. (2017). Comparative impact study of the European Social Survey (ESS) ERIC. Retrieved from https://www.europeansocialsurvey.org/docs/findings/ESS-Impact-study-Final-report.pdf

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come! Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods*, *15*, 722–752.

Levy, R. (2011). Bayesian data-model fit assessment for structural equation modeling. *Structural Equation Modeling*, *18*, 663–685.

Levy, R. & Choi, J. (2013). Bayesian structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling. a second course* (2nd ed., pp. 563–624). Charlottesville, NC: Information Age Publishing.

Meitinger, K. (2017). Necessary but insufficient: why measurement invariance tests need online probing as a complementary tool. *Public Opinion Quarterly*, *81*(2), 447–472.

Meredith, W. (1993). Measurement invariance, factor-analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543.

Meuleman, B. & Billiet, J. (2012). Measuring attitudes toward immigration in Europe: the cross-cultural validity of the ESS immigration scales. *ASK Research & Methods*, *21*(1), 5–29.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York: Routledge.

Munck, I., Barber, C., & Torney-Purta, J. (Online first). Measurement invariance in comparing attitudes towards immigrants among youth across Europe in 1999 and 2009: the alignment method applied to IEA CIVED and ICCS. *Sociological Methods & Research*.

Muthén, B. O. & Asparouhov, T. (2012). Bayesian SEM: a more flexible representation of substantive theory. *Psychological Methods*, *17*, 313–335.

Muthén, B. O. & Asparouhov, T. (2013). BSEM measurement invariance analysis. Mplus web notes, 17. Retrieved from https://www.statmodel.com/examples/webnotes/webnote17.pdf

Muthén, L. & Muthén, B. O. (2015). Mplus user's guide. Los Angeles, CA: Muthén & Muthén.

Oberski, D. L. (2014). Evaluating sensitivity of parameters of interest to measurement invariance in latent variable models. *Political Analysis*, *22*(1), 45–60.

Preston, I., Bauer, T., Card, D., Dustmann, C., & Nazroo, J. (2001). European Social Survey round 1 module proposal. Proposal for a module on immigration and attitudes. Retrieved from https://www.europeansocialsurvey.org/docs/round1/questionnaire/ESS1%5C_preston%5C_proposal.pdf

Saris, W. E., Satorra, A., & van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling*, *16*(4), 561–582.

Sarrasin, O., Green, E. G. T., Berchtold, A., & Davidov, E. (2012). Measurement equivalence across subnational groups: an analysis of the conception of nationhood in Switzerland. *International Journal of Public Opinion Research*, *25*(4), 522–534.

Schumacker, R. E. & Lomax, R. G. (2010). *A beginner's guide to structural equation modeling* (3rd ed.). New York: Routledge.

Siegers, P. & Davidov, E. (2010). Comparing basic human values in east and west germany. In T. Beckers, K. Birkelbach, J. Hagenah, & U. Rosar (Eds.), *Komparative empirische Sozialforschung [comparative empirical social research]* (pp. 43–63). Wiesbaden: VS Verlag.

Steenkamp, J.-B. E. M. & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*(1), 78–90.

Steinmetz, H., Schmidt, P., Tina-Booh, A., Wieczorek, S., & Schwartz, S. H. (2010). Testing measurement invariance using multigroup CFA: differences between educational groups in human values measurement. *Quality & Quantity*, *43*(4), 599–616.

Van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthén, B. O. (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), 1–15.

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods*, *5*(2), 139–158.

Vandenberg, R. J. & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70.

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 209–231). New York: Guilford Press.

Zercher, F., Schmidt, P., Cieciuch, J., & Davidov, E. (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), 1–11.