International Journal of Public Opinion Research Vol. 22 No. 4

doi:10.1093/ijpor/edq030 Advance Access publication 28 September 2010

How Harmful are Survey Translations? A Test with Schwartz's Human Values Instrument

Eldad Davidov[1,*] and Alain De Beuckelaer[2,*]

[1]University of Zurich, Switzerland, [2]Radboud University Nijmegen, Institute for Management Research, The Netherlands and Ghent University, Belgium

Abstract

One major challenge in international survey research is to ensure the equivalence of translated survey instruments across different cultures. In this study, we examine empirically the extent to which equivalence of survey instruments to measure human values can be established across cultures sharing the same language as opposed to cultures having a different language. We expect cultures using the same language to exhibit higher levels of equivalence. Our examination made use of a short (i.e., a 21-item) survey instrument to measure Schwartz's human values based on data from the second and the third rounds of the European Social Survey (ESS). The empirical results support our expectations.

One major challenge in cross-cultural comparative survey research is the necessity to ensure that multi-item survey instruments exhibit a high degree of *equivalence* across the cultural groups involved in the comparison (Smith, 2003). Essentially, equivalence of survey questions across cultural groups means that members of these different groups do not vary in terms of their interpretation of these survey questions and the way they use the scale (in some situations this statement may be too strong, and instead of a similar interpretation of the items equivalence would simply mean similar psychometric measurement properties). In a more technical sense, Horn and McArdle (1992) define measurement equivalence as "whether or not, under different conditions of observing and studying a phenomenon, measurement operations

*The two authors contributed equally to this work.
Correspondence concerning this article should be addressed to Alain De Beuckelaer, Radboud University Nijmegen, Thomas van Aquinostraat 1, P.O. Box 9108, 6500 HK Nijmegen, The Netherlands. E-mail: a.debeuckelaer@fm.ru.nl

yield measures of the same attribute" (p. 117). In very broad terms, the concept of equivalence may be described as a high degree of similarity in terms of psychometric or measurement properties of the survey instrument as observed in each cultural group (country or part of a country; see Schaffer & Riordan, 2003) under study.

Several authors have demonstrated that the establishment of equivalence across cultures is necessary before any meaningful cross-cultural comparisons may be conducted (Billiet, 2003; De Beuckelaer, 2005; Steenkamp & Baumgartner, 1998; Van den Berg, 2002; Van den Berg & Lance, 2000). However, efforts to guarantee equivalence of survey instruments may fail as questions are not always similarly understood across cultures, and the use of the scale of an instrument may be conditioned on the cultural context. When a different language is used across cultures, equivalence of the survey instrument is more likely to be absent, thus preventing meaningful cross-cultural comparisons (Tourangeau, Rips, & Rasinski, 2000). This problem may be especially evident in the measurement of affective survey items such as attitudes, opinions, normative beliefs, and values (Peytcheva, 2008).

Values are central to public discourse and are often viewed as deeply rooted, abstract motivations that guide, justify, and explain attitudes, norms, opinions, and actions (Feldman, 2003; Halman & De Moor, 1994; Rokeach, 1973; Schwartz, 1992). In this study we subject the human value theory (Schwartz, 1992) to strict tests of equivalence across cultures sharing the same language as opposed to cultures having a different language. We use data from two rounds of the European Social Survey (ESS),[1] which chose to measure the values in this theory by including a short 21-item instrument in its biannual studies. Before presenting the empirical results, we briefly review previous theoretical and empirical research on the establishment of equivalence with the same and different languages, and describe the theory and the ESS instrument. As expected, we find higher levels of equivalence of the value items among subjects from countries using the same language to complete the survey as compared to subjects from countries using a different language.

Previous Research

Theoretical Considerations

The establishment of a high degree of equivalence of translated survey instruments is contingent upon: (1) differences in languages as used in the study,

[1]The European Social Survey (ESS) project comprises a major collaborative effort designed to pioneer and validate a standard of methodology for cross-country surveys (e.g., using comparable modes of data collection). The project was one of the five winners of the prestigious '2005 EU Descartes Price'. (Retrieved from http://ec.europa.eu/research/press/2005/pdf/pr02122005_annex_winners_dp_research2005_en.pdf)

and (2) the cultural appropriateness of the translated survey instruments. According to Weech-Maldonado, Weidmer, Morales, and Hays (2001), a culturally appropriate translated survey instrument is (a) conceptually equivalent (i.e., equivalent in meaning and content), (b) technically equivalent to the source language (i.e., equivalent in grammar and syntax), (c) linguistically appropriate for the target population (i.e., readable and comprehensible), and (d) culturally competent (i.e., adequately reflecting cultural assumptions, norms, values, and expectations of the target population). Base on this definition it follows from here that when the same language is used across cultures, then equivalence of survey instruments is more likely.

More than fifty years ago, researchers in the fields of psychology and linguistics already introduced the idea that cultural differences in thought processes (cognition) are evident and interrelated with linguistic differences (Whorf, 1956). As thought processes are known to play a dominant role in the survey response process (Tourangeau et al., 2000), one may reasonably assume that language of survey administration may be partly responsible for cross-cultural/cross-country differences (bias) in survey results. According to Peytcheva (2008), the language of survey administration may affect all four stages of Tourangeau et al.'s survey response process, namely: (1) question comprehension (attending to the question and the instructions given, interpretation of key terms used in the question, and deciding what information to search for); (2) retrieval (activating and bringing information to mind from memory); (3) judgment (evaluating the information retrieved and integrating this information into an overall judgment); and (4) response. The response stage is comprised of an editing and a mapping phase. Editing refers to one's judgment evaluation before disclosing it, whereas mapping refers to the translation of the judgment into the format required in the survey questionnaire (e.g., choosing a particular response category to indicate one's agreement with a statement made). Especially survey items that differ across cultures or countries in terms of their affective characteristics (e.g., in terms of item sensitivity and proneness to social desirability) are expected to be prone to the biasing effect of language (Peytcheva, 2008, p. 2). Such items typically measure attitudes, normative beliefs (see also Berry & Sam, 1996), or human values (which is the focus of this study).

This may explain why survey methodologists have worked on issues related to survey translations (Acquadro, Jambon, Ellis, & Marquis, 1996; Harkness, 2003; McKay et al., 1996), and have worked on the development of good practice guidelines (Hambleton, 2001; Hambleton & Patsula, 1998; Van de Vijver & Hambleton, 1996; Weidmer, 2000) to ensure that both the survey translations and the scales used to answer the individual survey questions are maximally comparable.

The key question, however, is "how difficult is it to guarantee a sufficient level of measurement equivalence across cultures, when different languages are used to survey different cultures/countries?" Before we discuss some existing international studies that have examined measurement equivalence of (translated) survey instruments across a large number of societal cultures or countries, we first explain the level of measurement equivalence of survey instruments that is generally considered to be sufficient to make meaningful comparisons across cultures.

Testing for Measurement Equivalence Across Cultures

Once cross-cultural or cross-country data have been collected, researchers should assess whether the survey instruments used to measure the theoretical concepts under study exhibit measurement equivalence across cultures. As demonstrated by numerous authors (Billiet, 2003; De Beuckelaer, 2005; Steenkamp & Baumgartner, 1998; Van den Berg, 2002), failure to establish measurement equivalence across cultural groups may lead to erroneous conclusions regarding cross-cultural differences in concept means and the nature and the strength of empirical relations between the concepts studied. Provided that concept indicators may be perceived as consequences rather than causes of the concept, several statistical tools are available to test for measurement equivalence of concepts. Such tools should be applied prior to making any cross-cultural comparisons using the data at hand. Later, we provide more details on how to apply one of these tools, namely, multigroup mean and covariance structure (MACS) analysis.

Depending on the type of cross-group comparison one wants to make, different levels of equivalence are required (Scholderer, Brunsø, & Grunert, 2004; Steenkamp & Baumgartner, 1998; Van den Berg & Lance, 2000; Van de Vijver & Leung, 1997, p. 144). For instance, comparisons across countries or other cultural groups that involve structural relations between certain variables (i.e., structural comparisons) require the survey instrument to exhibit metric equivalence across groups.

Metric equivalence is established whenever individual survey questions (items) have identical factor loadings (i.e., slopes between the latent variable and the corresponding items) in all groups under study. Metric equivalence (i.e., an equivalence model specifying equal factor loadings across groups) is supported if such a model fits the data well and does not result in a significant reduction of model fit when compared with a model that does not set any measurement parameters to be equivalent across groups. The latter model may be conceived as the least constrained model and is referred to as the configural equivalence model. Chen (2007) suggested modern diagnostic criteria which are especially suitable to test for measurement equivalence in large sample studies. Chen's diagnostic criteria include differences in global model fit

indices such as comparative fit index (CFI) and root mean square error of approximation (RMSEA). Minimal differences in these model fit indices between the models may support a more restrictive model (for a criticism on the use of chi-square difference tests with large samples, see Cheung and Rensvold, 2002). Metric equivalence is a necessary condition for the meaningfulness of formal tests on higher levels of equivalence.

If the researcher aims at statistically comparing countries in terms of the absolute (mean) score of theoretically relevant concepts (i.e., making level comparisons), a third and even higher level of equivalence, scalar equivalence of survey questions/items (across groups) is required. Scalar equivalence, which is also referred to as full score equivalence (Van de Vijver & Leung, 1997, p. 144), is established whenever individual survey questions/items that are measuring a particular theoretical concept have identical factor loadings *and* intercepts (i.e., also identical scale origins) in all groups involved in the comparison (De Beuckelaer, 2005; Meredith, 1993; Steenkamp & Baumgartner, 1998; Van den Berg & Lance, 2000). The scalar equivalence model is supported if the model fit is acceptable and the model fit indices (mainly CFI and RMSEA) are not substantially reduced compared to corresponding model-fit indices of the metric equivalence model (Chen, 2007).

Several authors have suggested that when full equivalence is not ensured by the data, one may fall back to partial equivalence. Partial equivalence requires that only two items per concept exhibit measurement equivalence (Byrne, Shavelson, & Muthén, 1989; Steenkamp & Baumgartner, 1998).

If neither factor loadings nor intercepts are found to be equal, but concepts are measured by the same survey questions/items across countries, the model is considered to exhibit configural equivalence (across countries) only. With configural equivalence, a meaningful comparison of structural relations or absolute (mean) scores of theoretically relevant concepts across countries may be problematic and may result in some inaccuracy and imprecision. Admittedly, the degree of imprecision due to the absence of higher levels of equivalence may be smaller than the parameter differences that would be observed if comparisons are done. Unfortunately, it may be difficult to make an adequate judgment on this unless we have formally tested the level of equivalence across groups.

From the explanation above it is clear that, in comparison to metric equivalence, scalar equivalence (across groups) is a much more stringent psychometric criterion to meet. The interrelationship between the type of equivalence required (across countries) and the nature of the cross-group/cross-country comparison implies that the lower levels of measurement equivalence across groups (e.g., metric equivalence) are often not sufficient to ensure meaningful (unbiased) comparisons across the groups under study. As many international survey-based studies aim at making cross-group comparisons of

the concepts under study, researchers should be aware of the necessity to integrate formal checks on measurement equivalence in general, and scalar equivalence (full score equivalence) in particular as part of their statistical analysis. By not doing so, level comparisons of concept mean scores across groups may be erroneous and, therefore, possibly misleading (Billiet, 2003).

Commonly used procedures to check for measurement equivalence of multi-item scales such as exploratory factor analysis with target rotation (Caprara, Barbaranelli, Bermúdez, Maslach, & Ruch; Chan, Ho, Leung, Chan, & Yung, 1999; Van de Vijver & Leung, 1997, pp. 90–99) and multi-group covariance-based structural equation modeling (Jöreskog, 1971; Van de Vijver & Leung, 1997, pp. 99–107) are not sufficiently adequate as they only deal with information on covariance structures but not with information on mean structures. However, MACS analysis (Sörbom, 1974, 1978), as well as 'differential item functioning' approaches based on item response theory provide an adequate means to test for scalar equivalence of survey instruments across groups (Raju, Lafitte, & Byrne, 2002; Reise, Widaman, & Pugh, 1993; Stark, Chernyshenko, & Drasgow, 2006).

Previous Empirical Research

Several studies (Hulin, 1987; Liu, Borg, & Spector, 2004; Ryan, Chan, Ployhart, & Slade, 1999) assessed the cross-cultural equivalence of (translated) survey instruments across a set of countries, including countries with a common language and countries with a different language. These studies used data from employee surveys (Job Descriptive Index or JDI, Hulin, 1987 and Smith, Kendall, & Hulin, 1969; 11-item employee attitude survey, Ryan et al., 1999; German Job Satisfaction Survey, Liu et al., 2004; see also Borg, 2000, 2003) conducted among employees from one multinational organization. These studies have shown that: (a) Relatively high levels of equivalence such as metric equivalence (bear in mind that none of these studies assessed scalar equivalence!) may be found across country samples of employees who responded using the same language, and (b) relatively low levels of equivalence are found across country samples of employees who responded using a different language.

Even though these studies are informative, they are not capable of providing generalizable results. First of all, two of them (Hulin, 1987; Ryan et al., 1999) involved only a very limited number of cultures (four in Ryan et al., 1999, six in Hulin, 1987). The study by Liu et al. (2004) did involve a large number of countries but the researchers did not examine the highest level of measurement equivalence, namely, scalar equivalence of the survey instrument across countries. As mentioned before, the establishment of scalar equivalence is critical whenever the researcher aims to compare the construct means (i.e., level comparisons) across the groups under study.

A few recent studies (Davidov, 2008, 2009; Davidov, Meuleman, Billiet, & Schmidt, 2008; Davidov, Schmidt, & Schwartz et al., 2008) have shown that, in general, scalar equivalence is rarely established when several countries are compared using translated survey instruments. The results of the study by De Beuckelaer, Lievens, and Swinnen (2007) suggest that it is far more difficult to establish scalar equivalence across countries which belong to a different 'language group' (e.g., Sweden and Poland) compared to countries belonging to the same language group (e.g., Germany and Austria). In their study, De Beuckelaer et al. *always* had to reject the model of scalar equivalence in favor of the model of metric equivalence whenever comparisons were made across groups of countries in which more than one *different* language had been used during survey administration. However, their study also showed that the model of scalar equivalence across countries could often be retained if comparisons were made across countries in which a *common* language was used to collect the data. More specifically, the study by De Beuckelaer et al., concluded that scalar equivalence was established across the following groups of countries (or parts of countries), namely: (a) English-speaking countries (Australia, United Kingdom, United States, Canada); (b) Dutch-speaking (parts of) countries (Dutch-speaking part of Belgium and The Netherlands); and (c) German-speaking (parts of) countries (German-speaking part of Switzerland, Germany). As argued above, these findings seemed to suggest that translating a survey instrument into another language to allow its use in another country may jeopardize its cross-country equivalence. One should, however, realize that the study by De Beuckelaer et al., relied only on employee samples from one multinational organization. As such, the results of this study are very unlikely to be generalizable to the wider population on the country level. Furthermore, this study made use of an ad-hoc survey measure of work climate implying that results may not generalize to domains other than work-climate-assessment surveys.

In this study we contribute to this research line by subjecting the human-values scale (Schwartz, 1992) to strict tests of equivalence. We test its measurement equivalence across cultures/countries with respondents who use the same and or different ('mixed') languages. Our assessment will include measurements at two points in time and will, therefore, allow testing for the stability of research findings over time. To reach this goal we utilize representative data from the European Social Survey (ESS; Jowell, Kaase, Fitzgerald, & Eva, 2007).

Our theoretical considerations and previous empirical findings lead us to expect higher levels of equivalence among subjects from different countries using the same language when compared to groups of people from different countries using a different language to complete the survey. Before beginning

with the empirical analysis, a brief overview of the human values theory and previous studies testing its measurements is provided.

The Structure of Human Values

Schwartz (1994) defines human values as "desirable, transsituational goals, varying in importance, that serve as guiding principles in people's lives" (p. 21). His value theory includes 10 basic values with distinct motivational goals building on common elements in earlier approaches (Inglehart, 1990; Rokeach, 1973). The values are: hedonism, stimulation, self-direction, security, universalism, benevolence, conformity, tradition, power, and achievement. Table 1 presents the 10 values and the basic motivations behind them. For example, the motivational goal of the power value is social status and prestige, with control or dominance over people and resources. The motivational goal of achievement is personal success through demonstrating competence according to social standards.

In addition, Schwartz's theory suggests a structural relation between the values. Some values may oppose each other but other values may be closely related to each other. In other words, actions pursued to realize one value may be congruent or opposed to actions pursued to realize other values. For example, pursuing self-direction values may conflict with pursuing tradition values. Independent thought and action-choosing may conflict with acceptance of the customs and ideas that traditional culture or religion provide.

Table 1
Definitions of the Motivational Types of Values in Terms of Their Core Goal

Power	Social status and prestige, control or dominance over people and resources
Achievement	Personal success through demonstrating competence according to social standards
Hedonism	Pleasure and sensuous gratification for oneself
Stimulation	Excitement, novelty, and challenge in life
Self-direction	Independent thought and action-choosing, creating, exploring
Universalism	Understanding, appreciation, tolerance, and protection for the welfare of all people and for nature
Benevolence	Preservation and enhancement of the welfare of people with whom one is in frequent personal contact
Tradition	Respect, commitment, and acceptance of the customs and ideas that traditional culture or religion provide the self
Conformity	Restraint of actions, inclinations, and impulses likely to upset or harm others and violate social expectations or norms
Security	Safety, harmony, and stability of society, of relationships, and of self

Source: Sagiv and Schwartz (1995).

The theory proposes that we distinguish between 10 values. However, it is also suggested that the values form a continuum at a more basic level because the motivational differences of values are continuous rather than discrete (Davidov, Schmidt et al., 2008). As a result, adjacent values which are theoretically distinct from each other often appear in empirical studies as a single value (e.g., tradition and conformity, universalism and benevolence, or power and achievement).

On a higher level, the theory places the values around two bipolar dimensions. The first dimension contrasts self-transcendence, which includes the values universalism and benevolence, with self-enhancement, that includes the values power and achievement. The other dimension contrasts conservation, which includes the values tradition, conformity, and security, with openness to change, where the values self-enhancement and stimulation are situated. Hedonism is placed between the dimensions self-enhancement and openness to change (Schwartz, 1992, 1994, 2005). Several scales have been proposed to measure these values of which the most recent one is included in the ESS. Our study employs these measurements.

Measuring Human Values in the ESS

The European Social Survey (ESS) is a biannual European cross-country survey; questions to measure the human values have been included since 2002. As such, the ESS allows researchers to conduct cross-cultural comparative studies of human values using representative country data collected through comparable modes of data collection. Translation into each native language followed rigorous procedures outlined in Harkness (2003) that are designed to guarantee culturally appropriate translated survey instruments (Weech-Maldonado et al., 2001). The ESS includes 21 value indicators to measure the 10 values postulated by the theory. Two value indicators are given for each value and, as an exception, three for universalism because of its broad content. This questionnaire is based on Schwartz's original 40-indicator Portrait Values Questionnaire (PVQ; Schwartz, 2005; Schwartz et al., 2001). However, Schwartz shortened his PVQ battery of value indicators to allow its inclusion in the ESS. The portraits used as value indicators are double barreled and gender matched with the respondent. Both Schwartz (2003) and Saris and Gallhofer (2007) have shown empirically that double-barreled statements do not harm the quality (including the validity) of the data. The questions describe a person, and the respondent is asked to evaluate the extent to which this person is or is not like him or her. For example, the statement "It is important to him to make his own decisions about what he does. He likes to be free to plan and not depend on others" describes a person for whom self-direction is important. Respondents are

asked to rate such descriptions on a 6-point rating scale ranging from 1 (*not like me at all*) to 6 (*very much like me*). Table 2 presents the value questions and their labels, grouped by type of value.

Research with the (new) 21-indicator instrument to measure values in the ESS found that only seven value types from the original 10 values postulated by the theory could be identified in most countries with data from the first (2002/2003; see Davidov, Schmidt et al., 2008) and second (2004/2005; see Davidov, 2008) ESS rounds. At least three pairs of values had to be unified because they were strongly interdependent: power with achievement, universalism with benevolence, and tradition with conformity. Values that had to be unified are adjacent to each other in the circular structure that is the underlying (theoretical) continuum describing human values. Therefore, unifying them did not contradict theory; it rather suggested that there are not enough items in the ESS to measure 10 values (Davidov, Schmidt et al., 2008). Knoppen and Saris (2009) suggest that it is also a result of absent convergent and discriminant validity (Campbell & Fiske, 1959). All values exhibited metric equivalence across countries with data of the first ESS round. However, in the second ESS round, only 14 countries (the same 14 countries where seven values could be identified) exhibited metric equivalence. In addition, Davidov, Schmidt et al. (2008) and Davidov (2008) have introduced five paths (cross-loadings): two between the unified factor universalism–benevolence and the items important to be rich and important to have adventures; a third one between the unified factor conformity–tradition and the item important to get respect from others; a fourth one between the unified factor power–achievement and the item important to be modest; and a fifth between the unified value conformity–tradition and the item important to be rich.

In the empirical part we will extend this test by providing results for the level of equivalence of the human values across 'same-language' and 'different-language' countries using the same data. The countries that are included in the analysis with their respective sample sizes are presented in Table 3. Details on data collection techniques and response rates in each country are documented on the website http://www.europeansocialsurvey.org. The data for the analysis were taken from the website http://ess.nsd.uib.no.

We will test for equivalence across the following (subsets of) countries in ESS rounds 2 and 3, for reasons of convenience simply referred to as 'countries' in the following: Great Britain and Ireland (surveyed in English); France, the French-speaking part of Belgium, and the French-speaking part of Switzerland (all surveyed in French); Germany, Austria, and the German-speaking part of Switzerland (all surveyed in German); and The Netherlands and the Dutch-speaking part of Belgium (surveyed in Dutch).

Table 2
The ESS Human Values Scale [N(ESS Round 2) = 16,915; N(ESS Round 3) = 16,992]

Human value	Item number (according to its order in the ESS questionnaire) and wording (male version)
Self-Direction (SD)	1. Thinking up new ideas and being creative is important to him. He likes to do things in his own original way (ipcrtiv). 11. It is important to him to make his own decisions about what he does. He likes to be free to plan and not depend on others (impfree).
Universalism (UN)	3. He thinks it is important that every person in the world be treated equally. He believes everyone should have equal opportunities in life (ipeqopt). 8. It is important to him to listen to people who are different from him. Even when he disagrees with them, he still wants to understand them (ipudrst). 19. He strongly believes that people should care for nature. Looking after the environment is important to him (impenv).
Benevolence (BE)	12. It's very important to him to help the people around him. He wants to care for their well-being (iphlppl). 18. It is important to him to be loyal to his friends. He wants to devote himself to people close to him (iplylfr).
Tradition (TR)	9. It is important to him to be humble and modest. He tries not to draw attention to himself (ipmodst). 20. Tradition is important to him. He tries to follow the customs handed down by his religion or his family (imptrad).
Conformity (CO)	7. He believes that people should do what they're told. He thinks people should follow rules at all times, even when no-one is watching (ipfrule). 16. It is important to him always to behave properly. He wants to avoid doing anything people would say is wrong (ipbhprp).
Security (SEC)	5. It is important to him to live in secure surroundings. He avoids anything that might endanger his safety (impsafe). 14. It is important to him that the government insures his safety against all threats. He wants the state to be strong so it can defend its citizens (ipstrgv).
Power (PO)	2. It is important to him to be rich. He wants to have a lot of money and expensive things (imprich). 17. It is important to him to get respect from others. He wants people to do what he says (iprspot).
Achievement (AC)	4. It's important to him to show his abilities. He wants people to admire what he does (ipshabt). 13. Being very successful is important to him. He hopes people will recognize his achievements (ipsuces).
Hedonism (HE)	10. Having a good time is important to him. He likes to "spoil" himself (ipgdtim). 21. He seeks every chance he can to have fun. It is important to him to do things that give him pleasure (impfun).
Stimulation (ST)	6. He likes surprises and is always looking for new things to do. He thinks it is important to do lots of different things in life (impdiff). 15. He looks for adventures and likes to take risks. He wants to have an exciting life (ipadvnt).

Source: Davidov (2008).

Table 3
Sample Size by Country or Language Group in the Country and ESS Round Number

Country	ESS round 2 (2004–2005)	ESS round 3 (2006–2007)
1. Austria	2,256	2,405
2. Belgium (French-speaking part)	759	681
3. Belgium (Dutch-speaking part)	1,019	1,117
4. France	1,806	1,986
5. Germany	2,870	2,916
6. Great Britain	1,897	2,394
7. Ireland	2,286	1,800
8. The Netherlands	1,881	1,889
9. Switzerland (German-speaking part)	1,549	1,326
10. Switzerland (French-speaking part)	498	409
Total *N*	16,915	16,992

Results

Single-Country Analyses

Before testing the equivalence of the values across countries, we first tested models assessing the measurement of human values in each country separately. Byrne (2001, pp. 175–176) has acknowledged the importance of conducting single-country confirmatory factor analyses (CFAs; Bollen, 1989) prior to conducting ordinary multigroup confirmatory factor analyses (MGCFA) and MACS. At first, variance-covariance matrices were constructed as input for the models. Ten variance-covariance matrices—using Pearson correlations—were constructed for the 10 countries in ESS Round 2. Another 10 variance-covariance matrices were constructed for the 10 countries in ESS Round 3. We estimated all the subsequent models using the Amos 16.0 software program (Arbuckle, 2005). In all analyses, the maximum likelihood (ML) estimator was used. De Beuckelaer and Swinnen (2011) have shown that with such a large sample size, the use of ML and assuming normally distributed, continuous data produces consistent results with a model that accounts for ordinality. Table 4 provides the results of the single-country tests.

Results of the CFAs in each country show that it was not possible to identify all of the 10 values postulated by the theory in any of the countries with the ESS data. Some values were too strongly related and, therefore, needed to be unified. Column 2 of Table 4 reports how many values could be identified in each country. In general, six or seven values could be identified. In ESS Round 2, seven values were identified in Austria, the French- and Dutch- speaking parts of Belgium, France, Germany, The Netherlands, and the German-speaking part of Switzerland. Six values were identified in

Table 4
Single Country Analyses in ESS Rounds 2 and 3: Number of Values Identified and the Unified Values by Country

Country	Number of values identified	The unified values
ESS round 2		UNBE, COTR, POAC
1. Austria	7	UNBE, COTR, POAC
2. Belgium (French-speaking part)	7	UNBE, COTR, POAC
3. Belgium (Dutch-speaking part)	7	UNBE, COTR, POAC
4. France	7	UNBE, COTR, POAC
5. Germany	7	UNBE, COTR, POAC
6. Great Britain	6	UNBE, COTR, POAC, STSD
7. Ireland	6	UNBE, COTR, POAC, STSD
8. The Netherlands	7	UNBE, COTR, POAC
9. Switzerland (German-speaking part)	7	UNBE, COTR, POAC
10. Switzerland (French-speaking part)	6	UNBE, COTR, POAC, SECCOTR
ESS round 3		
11. Austria	7	UNBE, COTR, POAC
12. Belgium (French-speaking part)	6	UNBE, COTR, POAC, STSD
13. Belgium (Dutch-speaking part)	7	UNBE, COTR, POAC
14. France	7	UNBE, COTR, POAC
15. Germany	6	UNBE, COTR, POAC, STSD
16. Great Britain	7	UNBE, COTR, POAC
17. Ireland	7	UNBE, COTR, POAC
18. The Netherlands	6	UNBE, COTR, POAC, STSD
19. Switzerland (German-speaking part)	6	UNBE, COTR, POAC, STSD
20. Switzerland (French-speaking part)	7	UNBE, COTR, POAC

Notes: Three pairs of values are unified for all countries: Universalism with benevolence (UNBE), conformity with tradition (COTR), and power with achievement (POAC). Also, in this column, additional unified values are reported: Stimulation with Self-Direction (STSD), and security with conformity and tradition (SECCOTR).

Great Britain, Ireland, and the French-speaking part of Switzerland. In ESS Round 3, seven values were identified in Austria, the Dutch-speaking part of Belgium, France, Great Britain, Ireland, and the French-speaking part of Switzerland. Six values were identified in the French-speaking part of Belgium, Germany, The Netherlands, and the German-speaking part of Switzerland. Column 3 reports the values that had to be unified because they were too closely related. Results are consistent with findings in previous studies described in the last section (Davidov, 2008, Davidov, Schmidt et al., 2008) and suggest that the ESS presumably does not offer enough value indicators to distinguish between each of the single values (see also Schwartz & Boehnke, 2004). Knoppen and Saris (2009) suggest another reason for the requirement to unify values: The ESS value measurements do not possess discriminant validity (Campbell & Fiske, 1959), and—as a result—single values are too closely related to be modeled separately.

Cross-Country Comparisons with Different-Language Countries

In the cross-country equivalence analyses we follow procedural guidelines suggested by several authors (Cheung & Rensvold, 2002; De Beuckelaer, 2005; Steenkamp & Baumgartner, 1998; Van den Berg, 2002; Van den Berg & Lance, 2000). They describe two strategies to test for equivalence. The first is the 'bottom-up' strategy. According to this strategy, one increases the number of equality constraints (starting with configural equivalence, then metric, then scalar equivalence) until the model is not supported by the data. According to the second, 'top-down' strategy, one starts with the most constrained model (i.e., scalar equivalence) and releases equality constraints until the model is sufficiently supported by the data. For the current study we decided to implement the bottom-up strategy to inquire whether even weak forms of equivalence are absent.

First, a multigroup analysis with 10 countries was conducted twice: for ESS Round 2 data and for ESS Round 3 data. These analyses will enable us to make a rough estimate of the extent to which the value measurements are equivalent across different-language countries with some of those having the same and others having used different languages to complete the survey. The model used for the test is the same one that was confirmed for 20 countries in ESS Round 1 and for 14 countries in ESS Round 2 (Davidov, 2008; Davidov, Schmidt et al., 2008). This model included the seven values and five cross-loadings as reported in the previous section. The unified values in this model are universalism-benevolence, tradition-conformity, and power-achievement. The results are reported in Table 5.

The multigroup analysis in ESS Round 2 required unifying two additional pairs of values because they were related to each other too strongly and could not be modeled separately: between stimulation and self-direction, and between security and the unified value conformity-tradition. Thus, the 10 countries in ESS Round 2 did not provide support for the seven-value solution from the previous round.

The multigroup analysis in ESS Round 3 required unifying only one additional pair of values: stimulation and self-direction. Thus, also in the third round, the data from the 10 countries did not provide support for the seven-value solution.

The model fit in ESS rounds 2 and 3 was acceptable as can be seen in the fit measures reported from the third row onward in Table 5. The CFI value was higher than .90 and the RMSEA value was lower than .05. These fit measures were proposed by different authors to discern between models with a well-versus-poor fit to the data (Hu & Bentler, 1999; Marsh, Hau, & Wen, 2004). In other words, all 10 countries exhibited configural equivalence.

Next, we discuss the results of testing for metric equivalence across countries in rounds 2 and 3. For this purpose, we constrained the factor loadings of

Table 5
Global Fit Measures of Multigroup MACS models: Configural, Metric, and Scalar Equivalence Tests across 10 Cultural Units in ESS Rounds 2 and 3

Model	Required modifications	CFI	RMSEA	Pclose	Chi square	*df*	AIC	BCC
ESS Round 2								
Configural equivalence	ST and SD unified; SEC and COTR unified; Er(ipstrgv) ↔Er(impsafe); ST→ipstrgv	0.903	0.016	1.00	9,308	1,720	10,908	10,937
Metric equivalence		0.890	0.016	1.00	10,522	1,918	11,726	11,747
Partial metric equivalence		0.900	0.016	1.00	9,568	1,765	11,078	11,105
Partial scalar equivalence		0.853	0.019	1.00	13,287	1,810	14,707	14,732
ESS Round 3								
Configural equivalence	ST and SD unified; UNBE→impfree; Er(impsafe) ↔Er(ipadvnt)	0.906	0.017	1.00	9,501	1,670	11,201	11,234
Metric equivalence		0.893	0.017	1.00	10,766	1,859	12,088	12,114
Partial metric equivalence		0.903	0.017	1.00	9,798	1,724	11,390	11,420
Partial scalar equivalence		0.854	0.020	1.00	13,941	1,778	15,425	15,454

Notes: CFI: comparative fit index; RMSEA: root mean square error of approximation; SRMR: standardized root mean square residual; PCLOSE: probability of close fit; AIC: Akaike's information criterion; BCC: the Browne-Cudeck criterion; *df* : degrees of freedom. For a full description of the abbreviations of values and value indicators, see Table 2. If not otherwise indicated in column 2, the model in the test is the same model tested in Davidov, Schmidt et al. (2008) in the cross-country analyses with seven values (HE, ST, SD, SEC, and the unified values UNBE, POAC, and COTR) and five cross-loadings. Rightward arrow signifies that a modification requires estimating a cross-loading between a value and an indicator. Double-sided arrow signifies that a modification requires estimating the covariance.

the value indicators to be equal across the 10 countries. The global fit measures displayed in Table 5 do not support (full) metric equivalence for both rounds. Although the RMSEA is within the recommended criteria, the CFI falls below .90. Also, the difference in CFI between the configural equivalence and the metric equivalence models was above the recommended criteria (Chen, 2007). However, as we mentioned earlier, several authors have suggested that when full equivalence is not guaranteed, one may fall back to partial equivalence. In this context, partial equivalence requires that only two indicators per value have measurement parameters satisfying the required equivalence constraints (see Byrne et al., 1989; Steenkamp & Baumgartner, 1998). As Table 5 demonstrates, partial metric equivalence is supported by the data. The differences in the CFI and RMSEA fit measures between the configural and partial metric equivalence models were below the recommended criteria (Chen, 2007). Thus, one may conclude that the samples exhibit partial metric equivalence in ESS rounds 2 and 3. Hence, the determination of partial metric equivalence allows a comparison of the values' correlates (i.e., one particular type of structural comparison) among the 10 countries being analyzed.

Finally, we tested for scalar equivalence of value items across countries. For this test, data were augmented with information about the mean level of the indicators (mean and covariance structure analysis or MACS modeling, see Sörbom, 1974, 1978). In addition to factor loadings, the intercepts of value indicators across the countries were constrained to be the same. This test resulted in an unacceptable global fit for both rounds as can be seen by the indicators reported in Table 5, suggesting that one should reject the scalar equivalence model. Failure of the model to meet the scalar equivalence test implies that the value means may not be meaningfully compared across these countries. In other words, level comparisons across countries may be problematic.

Releasing parameters and constraining the parameters of only two indicators per value to be the same across countries (to find whether partial scalar equivalence may be supported by the data) did not result in any significant improvement of the model fit. To summarize, we found that neither full nor partial scalar equivalence were supported by the data.

Saris, Satorra, and Van der Veld (2009) propose an alternative method to detect model misspecification and evaluate model fit. They argued that model fit criteria do not provide an adequate indication about the size of the misspecification in the model. As a solution, they suggest using modification indices in combination with the expected parameter change (EPC) and the power of the modification index test. To enable researchers to use their approach, Van der Veld, Saris, and Satorra (2008) have developed a software program called Jrule or judgment rule (which works with output produced by

the LISREL software; see Van der Veld & Saris, 2011). An alternative version of this software, which works with Mplus software (see Muthén & Muthén, 2007), was developed by Oberski (2009; to download the software program, consult the website http://wiki.github.com/daob/JruleMplus). This alternative program along with cut-off criteria for misspecifications suggested by Saris et al. (2009) were applied in this study. In particular, a deviation of .10 or higher was suggested as a critical deviation for an item intercept. From here on we report the main results, but do not provide all the findings as this would seriously lengthen the manuscript. Outputs may be provided by the authors upon request.

Based on the program Jrule, full scalar and metric equivalence were rejected by the data across the 10 countries in rounds 2 and 3. Misspecifications were observed for different parameters and for all the values and countries in both rounds. However, similar to our prior results, partial metric equivalence was supported by the data.

It may be argued that equivalence is more difficult to achieve when the number of groups in the analysis is large. It could be the case that the relatively low levels of equivalence evidenced in the data are due to the fact that, in our earlier analyses, 10 countries or parts of countries were compared. To address this issue, 30 random dyads (15 for each round) of different-language countries were drawn and their level of equivalence was tested with the two methods of analysis. It turned out that all pairs of countries reached partial scalar equivalence. However, partial scalar equivalence was *not* reached for all the values in the analysis. More specifically, in 73% of the pairs (22 out of 30), at least one value (mostly hedonism but often also self-direction, and the unified values conformity-tradition and power-achievement) did not achieve partial scalar equivalence. In 33% of the pairs (10 out of 30), at least two values did not reach partial scalar equivalence. Finally, in 20% of the pairs (6 out of 30), 3 values or more did not reach partial scalar equivalence. Conclusions were consistent using the differences in global fit measures across alternative models method and the Jrule method. Next, we turn to the analysis of pairs of same-language countries or groups of countries. This will allow us to compare the levels of equivalence reached when the same language is used and to examine which particular pairs of countries reach higher levels of equivalence.

Cross-Country Comparisons with Same-Language Countries

In this phase we performed multiple group comparisons across pairs of countries using the same language in each round. As previously argued, here we expect to find higher levels of equivalence compared with the test across the different-language countries. We performed 16 multiple-group comparisons across pairs of same-language countries: eight of them were conducted using

data from ESS Round 2, and included comparisons between Germany and Austria, Germany and the German-speaking part of Switzerland, Austria and the German-speaking part of Switzerland, France and the French-speaking part of Belgium, France and the French-speaking part of Switzerland, the French-speaking part of Belgium and the French-speaking part of Switzerland, Great Britain and Ireland, and the Netherlands and the Dutch-speaking part of Belgium. Another eight comparisons (between the same groups) were conducted using data from ESS Round 3. A detailed report of the global fit measures and the misspecifications are available from the authors upon request.

Some model modifications were required in the multiple-group comparisons. These modifications were in line with previous research (Davidov, 2008) and included either unifying an additional pair of adjacent values, a few cross-loadings, or releasing error correlations.

It turned out that all pairs of countries reached partial scalar equivalence in the multigroup analyses. However, partial scalar equivalence was *not* reached for all the values in the analysis. In 50% of the pairs (8), at least one value did not achieve partial scalar equivalence. In 13% of the pairs (2), two values did not reach partial scalar equivalence. Results were consistent with the two methods of analysis (i.e., using global fit measures and using Jrule) and across both rounds. This finding is much better than that for the different-language pairs (50% vs. 73% reported for different-language countries; 13% vs. 33% reported for different-language countries; and 0% vs. 20% reported for different-language countries) and indicates that, on average, significantly more values reached at least scalar equivalence. However, differences in the findings were evidenced across different languages. Dutch- and English-speaking countries always exhibited partial scalar equivalence for all human values and ESS rounds. French- and German-speaking countries almost never exhibited partial scalar equivalence for all values. It may indicate a larger cultural or linguistic distance among French- and German-speaking countries/parts of countries when we compare this to Dutch-speaking or English-speaking countries in Europe. However, more units of analysis are necessary to test this proposition.

Discussion

In this study we assessed to what extent translations may harm the cross-country equivalence of Schwartz's 21-item human values instrument as implemented in the ESS. Measurement equivalence tests were conducted across groups of countries using the same language during survey administration and groups of countries using a different language during survey administration. The results of our analyses were generally consistent across

two rounds of the ESS, strengthening our confidence in the temporal stability of our study results.

The empirical findings supported our expectation that higher levels of measurement equivalence were to be found across countries sharing the same language. In particular, very high levels (i.e., partial scalar or scalar) of equivalence were found especially in both English-speaking and Dutch-speaking countries in both ESS Round 2 and Round 3. This finding did not surprise us that much, given that the earlier study by De Beuckelaer et al. (2007) also reported full scalar equivalence of survey measures across four English-speaking countries located in very different regions of the world (in particular: Australia, United Kingdom, Canada, and the United States). In very much the same way, the same study also reported (full) scalar equivalence of survey measures across countries having Dutch or German as their common language (i.e., the Dutch-speaking part of Belgium and The Netherlands, and the German-speaking part of Switzerland and Germany, respectively). Hence, measurement equivalence assessment across same-language countries as conducted in our study revealed patterns of measurement equivalence which had been observed in earlier empirical research which did not rely on cross-country representative samples and dealt with an ad-hoc measure of work climate.

Our study further showed that across pairs of countries with different languages, Schwartz's human values instrument exhibited partial scalar equivalence for significantly less values compared with the same-language country pairs. As such, lower levels of equivalence were obtained. This finding is also in line with De Beuckelaer et al. (2007) who had to reject the model of scalar equivalence in favor of the model of (full) metric equivalence each time multiple languages were involved.

In sum, the empirical findings from our study provide some empirical evidence to support the belief that translations, which are a necessity in most international survey research, may seriously distort the comparability (or measurement equivalence) of survey data across countries. This may apply even when rigorous translation procedures as implemented in the ESS (Harkness, 2003) are used. Realizing that: (a) despite the long-standing debate on the "Whorfian hypothesis" (see Hunt & Agnoli, 1991), cultural differences in thought processes are seen as being interrelated with intrinsic differences in languages (see also Whorf, 1956), and (b) thought processes influence all stages of the survey response process (Tourangeau et al., 2000), we expected an influence of translations on the comparability of data across countries. As mentioned above, the results of this study confirmed our expectations.

From a practical point of view, we would like to stress that establishing an adequate level of measurement equivalence is critical whenever a researcher aims to make cross-country comparisons. For this reason a researcher should

always formally check whether the level of measurement equivalence needed (i.e., partial metric equivalence for structural comparisons; partial scalar equivalence for level comparisons) is also supported by the data. Together, with some other studies, our study also showed that higher level of equivalence of (multi-item) survey measures was harder to establish for all values in the model, especially if multiple languages are used to administer the survey. Including more concept (i.e., human values) indicators in the survey may help to increase the chance of establishing higher levels of equivalence but including many concept indicators is often not realistic because of practical constraints (see Schwartz's human value scale in the ESS).

Even though our study showed that levels of measurement equivalence tend to be higher across same-language countries when compared to countries with mixed languages, we should interpret these findings with some caution. Due to the quasi-experimental nature of our research and the limited number of countries (or parts of countries) sharing the same language included in the ESS, it was not possible to provide adequate control for cultural distance, at least not for those aspects of culture which are not related to either language or the culturally determined aspects of one's thought processes when answering Schwartz's value survey (see the work by Whorf, 1956 and Peytcheva, 2008). So, we have no absolute certainty that the equivalence patterns found in the ESS data are entirely due to language; they may also be caused—at least to some extent—by "nonlanguage-related aspects of culture". Indeed, English-speaking and Dutch-speaking countries displayed partial scalar equivalence for all values but German-speaking and French-speaking countries reached partial scalar equivalence only for a subset of the values in the model. As we have previously mentioned, it may indicate a larger cultural or linguistic distance among French-speaking and German-speaking countries/parts of countries when we compare this to Dutch-speaking or English-speaking countries in Europe. However, more units of analysis are necessary to provide further empirical evidence for this proposition.

To assess the effect of language/translation over and above the effect of nonlanguage-related aspects of culture one would need similar data as those used in this study but from a larger number of culturally diverse countries within the same language group. We are not aware of the existence of such a data set, at least not one dealing with the measurement of human values. However, in one empirical study that we cited earlier (De Beuckelaer et al., 2007), the effect of culture over and above the effect of language was not found to be substantial in a sample which was based on a larger number of countries within several cultural clusters. Obviously, future work based on large, representative country samples is needed to evaluate whether the conclusions from this earlier study can be generalized to other domains of survey measures (e.g., Schwartz's human values) and other countries.

A further limitation of our study concerns its exclusive focus on (one instrument to measure) human values. Despite the fact that human values are central to public discourse today, and are often considered to be an important determinant of certain types of behavior, opinions, and attitudes, it would be worthwhile to conduct a similar equivalence study as this one using other key predictors of human behavior. For instance, one could think of the 'Big Five', that is, five universal personality traits (see Costa & McCrae, 1992). One of the major problems with conducting such a study concerns the requirement of large data files which are also representative for the different cultures/countries under study. In that sense, the European Social Survey has really provided us with a unique data set to study human values across a large number of countries with similar and different languages from all over Europe.

Acknowledgments

We are indebted to Johnny Fontaine (Ghent University, Belgium), Willem Saris (ESADA, Barcelona, Spain), and two anonymous reviewers for their valuable comments on a previous version of this paper. In addition, we would like to thank Lisa Trierweiler for the English proof of the manuscript. A previous version of this paper was presented at the ESRA conference, Warsaw, June 29 to July 3, 2009 and at the International Workshop on Comparative Survey Design and Implementation (CSDI), Lausanne, March 25–27, 2010.

References

Acquadro, C., Jambon, B., Ellis, D., & Marquis, P. (1996). Language and translation issues. In B. Spilker (Ed.), *Quality life and pharmacoeconomics in clinical trials* (2nd ed., pp. 75–82). Philadelphia: Lippincott-Raven.

Arbuckle, J. L. (2005). *Amos 6.0 user's guide*. Chicago, IL: SPSS Inc.

Berry, J., & Sam, D. (1996). Acculturation and adaptation. In J. Berry, M. Segall & C. Kagitcibasi (Eds.), *Handbook of cross-cultural psychology: Social behavior and applications* (pp. 291–325). Boston, MA: Allyn & Bacon.

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

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

Borg, I. (2000). *Führungsinstrument Mitarbeiterbefragung* [Management Instrument Employee Survey] (2nd ed.). Göttingen, Germany: Hogrefe.

Borg, I. (2003). *Führungsinstrumen Mitarbeiterbefragung: Theorien, Tools und Praxisehrfahrungen* [Management Instrument Employee Survey: Theories, tools, and practical experiences] (3rd ed.). Göttingen, Germany: Verlag für Angewandte Psychologie.

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

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

Campbell, D. T, & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait multimethod matrices. *Psychological Bulletin, 56*, 81–105.

Caprara, G. V., Barbaranelli, C., Bermúdez, J., Maslach, C., & Ruch, W. (2000). Multivariate methods for the comparison of factor structures in cross-cultural research: An illustration with the Big Five questionnaire. *Journal of Cross-Cultural Psychology, 31*, 437–464.

Chan, W, Ho, R. M., Leung, K., Chan, D. K.-S., & Yung, Y.-F. (1999). An alternative method for evaluating congruence coefficients with Procrustes rotation: A bootstrap procedure. *Psychological Methods, 4*, 378–402.

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

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

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Questionnaire Inventory (NEO-PI-RTM) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

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

Davidov, E. (2009). Measurement equivalence of nationalism and constructive patriotism in the ISSP 2003: 34 countries in comparative perspective. *Political Analysis, 17*, 64–82.

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

Davidov, E., Meuleman, B., Billiet, J., & Schmidt, P. (2008). Values and support for immigration. A cross-country comparison. *European Sociological Review, 24*, 583–599.

De Beuckelaer, A. (2005). *Measurement invariance issues in international management research*. Unpublished doctoral dissertation, Hasselt University, Diepenbeek, Belgium.

De Beuckelaer, A., & Swinnen, G. (2011). Biased latent variable mean comparisons due to measurement non-invariance: A simulation study. In E. Davidov, P. Schmidt & J. Billiet (Eds.), *Methods and applications in cross-cultural analysis* Taylor & Francis.

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

Feldman, S. (2003). Values, ideology, and structure of political attitudes. In D. O. Sears, L. Huddy & R. Jervis (Eds.), *Oxford handbook of political psychology* (pp. 477–508). New York, NY: Oxford University Press.

Halman, L., & De Moor, R. (1994). Value shift in Western societies. In P. Ester, L. Halman & R. de Moor (Eds.), *The individualizing society: Value change in Europe and North America* (pp. 1–20). Tilburg, The Netherlands: Tilburg University Press.

Hambleton, R. K. (2001). The next generation of the ITC test translation and adaptation guidelines. *European Journal of Psychological Assessment, 17*, 164–172.

Hambleton, R. K., & Patsula, L. (1998). Adapting tests for use in multiple languages and cultures. *Social Indicators Research, 45*, 153–171.

Harkness, J. A. (2003). Questionnaire translation. In J. A. Harkness, F. J. R. Van de Vijver & P. Ph. Mohler (Eds.), *Cross-cultural survey methods* (pp. 35–56). New York, NY: John Wiley.

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

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55.

Hulin, C. L. (1987). A psychometric theory of evaluations of item and scale translations: Fidelity across languages. *Journal of Cross-Cultural Psychology, 18*, 115–142.

Hunt, E., & Agnoli, F. (1991). The Whorfian hypothesis: A cognitive psychology perspective. *Psychological Review, 98*, 377–389.

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

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

Jowell, R., Kaase, M., Fitzgerald, R., & Eva, G. (2007). The European Social Survey as a measurement model. In R. Jowell, M. Kaase, R. Fitzgerald & G. Eva (Eds.), *Measuring attitudes cross-nationally* (pp. 1–29). London, UK: Sage.

Knoppen, D., & Saris, W. (2009). Do we have to combine values in the Schwartz' human values scale? A comment on the Davidov studies. *Survey Research Methods, 3*, 91–103.

Liu, C., Borg, I., & Spector, P. E. (2004). Measurement equivalence of the German job satisfaction survey used in a multinational organization: Implications of Schwartz's culture model. *Journal of Applied Psychology, 89*, 1070–1082.

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

McKay, R. B., Breslow, M. J., Sangster, R. L., Gabbard, S. M., Reynolds, R. W., Nakamoto, J. M., & Tarnai, J. (1996). Translating survey questionnaires: Lessons learned. *New Directions for Evaluation, 70*, 93–105.

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

Muthén, L. K., & Muthén, B. O. (2007). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.

Oberski, D. L. (2009). *Jrule for Mplus version 0.91* [computer software]. Retrieved from the Github Social Coding website: http://wiki.github.com/daob/JruleMplus/.

Peytcheva, E. A. (2008). *Language of administration as a source of measurement error: Implications for surveys of immigrants and cross-cultural survey research* Unpublished doctoral dissertation, University of Michigan.

Raju, N., Lafitte, L. J., & Byrne, B. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87*, 517–529.

Reise, S., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*, 552–566.

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

Ryan, A. M., Chan, D., Ployhart, R. E., & Slade, L. A. (1999). Employee attitude surveys in a multinational organization: Considering language and culture in assessing measurement equivalence. *Personnel Psychology, 52*, 37–58.

Sagiv, L., & Schwartz, S. H. (1995). Value priorities and readiness four out-group social contact. *Journal of Personality and Social Psychology, 69*, 437–448.

Saris, W. E., & Gallhofer, I. N. (2007). *Design, evaluation, and analysis of questionnaires for survey research*. New York, NY: Wiley.

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

Schaffer, B. S., & Riordan, C. (2003). A review of cross-cultural methodologies for organizational research: A best-practices approach. *Organizational Research Methods, 6*, 169–215.

Scholderer, J., Brunsø, K., & Grunert, K. (2004). Cross-cultural validity of the food-related lifestyles instrument (FRL) within Western Europe. *Appetite, 42*, 197–211.

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

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

Schwartz, S. H. (2003). A proposal for measuring value orientations across nations. ESS Questionnaire Development Report (chap. 7). Retrieved from: http://www.europeansocialsurvey.org.

Schwartz, S. H. (2005). Basic human values: Their content and structure across countries. In A. Tamayo & J. B. Porto (Eds.), *Valores e comportamento nas organizações* [Values and behavior in organizations] (pp. 21–55). Petrópolis, Brazil: Vozes.

Schwartz, S. H., & Boehnke, K. (2004). Evaluating the structure of human values with confirmatory factor analysis. *Journal of Research in Personality, 38*, 230–255.

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

Smith, P. C., Kendall, L., & Hulin, C. L. (1969). *The measurement of satisfaction in work and retirement*. Chicago, IL: Rand McNally.

Smith, T. W. (2003). Developing comparable questions in cross-national surveys. In J. Harkness, F. J. R. van de Vijver & P. Möhler (Eds.), *Cross-cultural survey methods* (pp. 69–92). New York, NY: Wiley.

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

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

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified approach. *Journal of Applied Psychology, 91*, 1292–1306.

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

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *Psychometric theories of survey response*. Cambridge: Cambridge University Press.

Van de Vijver, F. J. R., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist, 1*, 89–99.

Van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data-analysis for cross-cultural research*. Thousand Oaks, CA: Sage.

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

Van der Veld, W., Saris, W. E., & Satorra, A. (2008). *Jrule 2.0: User manual.* Unpublished manuscript (Internal report, Radboud University Nijmegen, The Netherlands).

Van den Berg, R. J. (2002). Towards a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5*, 139–158.

Van den Berg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4–69.

Weech-Maldonado, R., Weidmer, B. O., Morales, L. S., & Hays, R. D. (2001). Cross-cultural adaptation of survey instruments. The CAHPS experience. In N. Cynamon & R. Kulka (Eds.), *Proceedings of the Seventh Conference on Health Survey Research Methods*. Hyattsville, MD: DHHS.

Weidmer, B. (2000). *Designing and adapting health surveys for cross-cultural research.* Paper presented at the 24th Annual Meeting of the Society of General Internal Medicine, San Diego, CA.

Whorf, B. L. (1956). *Language, thought, and reality: Selected writings of Benjamin Lee Whorf*. New York, NY: Wiley.

Biographical Notes

Eldad Davidov is Professor of Sociology at the University of Zurich (CH). His research interests are applications of structural equation modeling to survey data, especially in cross-cultural and longitudinal research. Applications include human values, national identity, and attitudes toward immigrants and other minorities.

Alain De Beuckelaer is a tenured faculty member (“Universitair Docent”) at Radboud University Nijmegen (NL), and a Senior Researcher at Ghent University (BE) in the Department of Personnel Management, Work and Organizational Psychology and the Department of Macro- & Structural Sociology. His research includes international management research (e.g., human resource management, organizational behaviour, marketing), cross-cultural (survey) research methodology, and multivariate statistical methods.